OUTCOME MEASURES OF PHYSICAL FUNCTION IN ADULTS with a UNILATERAL LOWER LIMB AMPUTATION DURING PROSTHETIC REHABILITATION: USE IN CLINICAL PRACTICE AND PSYCHOMETRIC PROPERTIES

JUDY SCOPES

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

QUEEN MARGARET UNIVERSITY

2016

Abstract

The aim of this thesis is to inform clinicians and researchers of the reliability and responsiveness of the most commonly used outcome measures in prosthetic rehabilitation in the UK. In addition, this thesis supports the call for more studies of high methodological quality to provide evidence of the psychometric properties of outcome measures of physical function in lower limb amputees.

A survey (study I) of Allied Health Professionals established that the outcome measures used most often during prosthetic rehabilitation in the UK were: the Timed Up and Go (TUG), a timed walk test, the Locomotor Capability Index (LCI) and its modified version (LCI-5), the Socket Comfort Score (SCS) and the Special Interest Group in Amputee Medicine (SIGAM) Mobility Grades.

A standardised quality checklist (COSMIN) was used in a systematic review (study II) to measure the methodological quality and strength of evidence of the published literature that reported on the psychometric properties of outcome measures used to measure physical function during prosthetic rehabilitation. The review found mixed methodological quality ratings and many studies with small sample sizes rendering the strength of the evidence indeterminate. A limited number of studies commented on limits of agreement and measurement error when reporting on reliability. Even fewer studies reported on responsiveness with only one reporting minimally clinically important difference (MCID) values.

Values for consistency, agreement and measurement error, were calculated for the top five commonly used outcome measures as identified from the survey, using a test-retest study design with a period of 7 days between tests (study III). Minimum detectable change (MDC) values were calculated for the SIGAM, LCI-5, TUG and 2MWT. The EQ-5D-5L™, a measure of the global health of the respondent, was also included as knowledge of its psychometric properties in a population of pwLLA is unknown. However, reliability could not be confirmed for the EQ-5D-5L or the SCS in this population.

A longitudinal study (study IV), based during the early rehabilitation period (mean 84 days) following provision of a primary prosthesis, gathered data to calculate indices of responsiveness for the same six outcome measures. Effect sizes were presented for five measures: SIGAM, LCI-5, TUG, 2MWT, SCS and EQ-5D-5L. Minimal clinically important difference values were also presented for the first time for all the outcome measures in this population. A patient reported change questionnaire was used as the anchor in a Receiver Operator Characteristic (ROC) curve analysis to establish the MCID values.

Key Words: Lower limb amputation, Prosthetic rehabilitation, Outcome measurement, Reliability, Measurement error, Responsiveness.

Acknowledgements

I would like to take this opportunity to thank several individuals and groups of people who have helped and guided me to the completion of this thesis. Though I cannot name everyone, there are a few I would like to single out for a special mention.
Without the amputees who agreed to participate, there would be no studies, and I am grateful to all of them. But I would like to recognise all the amputees I have met and worked with throughout my career, who inspired me to ask the questions that sparked the idea for this PhD in the first place.

I would like to thank my supervising team of Dr Marietta van der Linden and Professor Nigel Gleeson who provided me with motivation, guidance and feedback throughout the past five years, but especially Marietta who was always there with her quick and insightful responses to my sometimes "stupid" questions.

The staff working in the SMART centre and Sutherland Ward at Astley Ainslie Hospital, Edinburgh welcomed me back warmly as an ex-colleague who had a mad idea to do some research. They helped me find and recruit my study participants, without whom there would be no thesis. They also supported me with kind words and coffee throughout the long recruitment period. Thank you especially to Catriona Mawdsley, Katy Bryce and the physiotherapy team at Astley Ainslie Hospital. I would also like to recognise the support and guidance that David Gow gave me; listening to my idea in the first place and guiding me through the early years of this PhD with his sage counsel.

I am grateful to the State of the Art Prosthetic Group at NHS Lothian and the charity PORTer from whom I received financial assistance to complete the survey and assistance with the study expenses associated with the participant studies.

Finally, I would like to thank my family and friends for sticking by me when I had the idea to go back to studying after so many years. I would not have made it without their love and support. Especially Nik, who has proof read nearly every word for me, and gone beyond the call of duty to make sure I was fed, clothed and watered throughout it all. Thank you for believing in me..

Table of Contents

List of Tables

List of Figures

List of Appendices

Abbreviations

Abbreviations used within the thesis

ACQ	Activity Change Questionnaire
ADL	Activities of Daily Living
AHP	Allied Health Professional
ANOVA	Analysis of variance
AUC	Area under the curve
BACPAR	British Association of Physiotherapists in Amputee Rehabilitation
BAPO	British Association of Prosthetists and Orthotists
BSRM	British Society of Rehabilitation Medicine
COMET	Core Outcome Measures in Effectiveness Trials
CONSORT	Consolidated Standards for Reporting Trials
COS	Core outcome set
COSMIN	COsensus-based Standards for the selection of health Measurement INstruments
CTT	Classic Test Theory
EQUATOR	Enhancing the QUAlity and Transparency Of health Records
ES	Effect size
EWA	Early walking aid
GRI	Guyatt Responsiveness Index
HR-PRO	Health related patient reported outcome
HR-QoL	Health related Quality of Life
IC	Internal Consistency
ICC	Intra-class correlation coefficient
ICF	International Classification of Function, disability and health
ISPO	International Society of Prosthetics and Orthotics
ITT	Item Test Theory
LoA	Limits of agreement
MCDC	Minimal clinically detectable change
MCID	Minimal clinically important change
MDC	Minimal detectable change
MDT	Multi-disciplinary team
MID	Minimal important difference
MPCK	Micro processor controlled knee
OMERACT	Outcome Measures in Rheumatology Clinical Trials
PAD	Peripheral arterial disease
PIS	Participant information sheet

PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-analyses
PROM	Patient reported outcome measure
PR-QoL	Prosthetic-related Quality of Life
pwLLA	Person/People with a Lower limb Amputation
ROC	Receiver Operator Curve
SD	Standard deviation
SDC	Smallest detectable change
SEM	Standard error of measurement
SIP	Study information pack
SPARG	Scottish Physiotherapy Amputee Research Group
SRM	Standardised response mean
STROBE	Strengthening the Reporting of Observed Studies in Epidemiology
SV(1)	Study Visit (1)
TFA	Trans-femoral amputation
TI(1)	Time Interval (1)
TREND	Transparent Reporting of Evaluations with Nonrandomised Designs
TTA	Trans-tibial amputation
TV(1)	Test Visit (1)
UNIPOD	United National Institute for Prosthetics & Orthotics Development
WHO	World Health Organisation

Abbreviations used for outcome measures within the thesis

2MWT	Two minute walk test
6MWT	Six minute walk test
AAS	Amputee Activity Score
ABC scale	Activities-specific Balance Confidence scale
ADAPT	Assessment of Daily Activity Performance in Trans-femoral amputees
AMP	Amputee Mobility Predictor
BBS	Berg Balance Scale
COPM	Canadian Occupational Performance Measure
FAI	Frenchay Activities Index
FIM	Functional Independence Measure
FMA	Functional Measure for Amputees
GAS	Goal Attainment Scale
HAI	Hill Assessment Index
LCI	Locomotor Capability Index
LCI-5	Locomotor Capability Index (modified)

NHP	Nottingham Health Profile
OLST	One Leg Stand Test
OPCS scale	Office of Population Censuses and Surveys scale
OPOT	Orthotics and Prosthetics National Outcomes Tool
OPUS	Orthotics and Prosthetics Users' Survey
PCI	Physiological Cost Index
PEQ	Prosthesis Evaluation Questionnaire
POGS	Prosthetic Observational Gait Score
PPA	Prosthetic Profile of the Amputee
PSFS	Patient-Specific Functional Scale
RMI	Rivermead Mobility Index
SCS	Socket Comfort Score
SIGAM	Special Interest Group in Amputee Medicine
SIP	Sickness Impact Profile
TAPES	Trinity Amputation and Prosthesis Experience Scale
TUG	Timed Up and Go
TWT	Timed Walk Test

Publications / Presentations arising from this thesis

Scopes J, Van der Linden M, Gleeson N, 2015. ISPO World Congress
Responsiveness of functional outcome measures used with lower limb amputees during early prosthetic rehabilitation [Platform Presentation] Lyon, 25th June 2015

Scopes J, Van der Linden M, Gleeson N, 2015. World Confederation of Physical Therapists
Psychometric properties of outcome measures of physical function used with lower limb amputees during prosthetic rehabilitation [Platform Presentation] Singapore, 3rd May 2015

Scopes J, Van der Linden M, Gleeson N, 2015. World Confederation of Physical Therapists
Minimal detectable change values of common outcome measures used in lower limb prosthetic rehabilitation in the UK [Poster Presentation] Singapore, 2nd May 2015

Scopes J. M. 2013 British Association of Chartered Physiotherapists in Amputee Rehabilitation (BACPAR) Annual Conference
A survey of health professionals and outcome measures for amputees [Platform Presentation] Wolverhampton, Nov 2013

Scopes J, Van der Linden M, 2013 Chartered Society of Physiotherapy Annual Conference
Important Factors and Outcome Measures used in Prosthetics Rehabilitation: A Survey of Health Professionals and Patients [Poster Presentation] Birmingham, Oct 2013

1 Chapter 1 Introduction

1.1 Introduction

A person's life is changed forever when a limb is amputated. Whether the amputation is performed to save their life following a traumatic accident; or follows a gradual deterioration in their clinical condition, amputation surgery will alter their pre-morbid physical and emotional state. A variety of factors influence an amputee's rehabilitation, and understanding the impact and context of these factors is essential to a successful outcome for the patient. However, the focus in this thesis will be on the physical changes that a person with a lower limb amputation (pwLLA) experiences, and how these changes may be measured, in particular following delivery of their first artificial limb or prosthesis.

It is accepted that the provision of a prosthesis following a lower limb amputation will change the physical performance of a pwLLA (Sansam et al. 2009, Ostler et al. 2014). However, arguments continue about the distinctions between different prosthetic components regarding their impact on this performance. In monetary terms the difference between two prosthetic knee joints can be thousands of pounds, but it is still a matter of debate whether differences in the functional abilities of a pwLLA due the use of two different joints can be detected. Is one knee "better" than another? What does "better" mean? If, for example, this means, "better mobility" then how is mobility measured? According to an Evidence Note produced in 2012 by the Scottish Government body Health Improvement Scotland (Health Improvement Scotland 2012), evidence suggests that "in healthy and active younger people, who have had a trans-femoral amputation the C-Leg® (Otto Bock, Duderstadt) may improve health outcomes (e.g. body image, safety, energy efficiency, gait and functionality) compared with mechanically controlled knees". However, the strongest evidence was from crossover randomised trials, and the authors of the evidence note commented that these studies were "generally small" (range 1-41) and "methodologically weak" (non-blinded, non-randomised and not generalisable to the wider population of pwLLA) (Highsmith et al. 2010, Theeven et al. 2011). An issue highlighted within this Highsmith review of prosthetic research was bias, as many studies included had also "received sponsorship from the manufacturers".

The price for the two knees described was given as £2,000 for the mechanical knee and £16,000 for the C-leg, a microprocessor knee (MPCK). The prices were correct at the time the evidence note was produced, though costs for new generation MPCKs have risen considerably since. Health Improvement Scotland go on to conclude that there was little evidence relating to older people with chronic illness or reduced function, and there is "insufficient evidence to determine whether the MCPKs are cost effective compared to the mechanical knees". With healthcare costs rising it is incumbent on the clinicians prescribing these different components, to understand the impact of one over the other.

The current condition of a patient can be captured by an outcome measure. If an outcome measure is repeated then a change in their condition may then be determined. Thus, data collected from outcome measures can provide the answer to the questions posed above, but only if the outcome measure used is fit for that purpose. For example, when looking at how mobile an amputee is around their home and outdoors using his / her prosthesis, data from the Wheelchair Skills Test (Kirby et al. 2004) will only provide information on how mobile an amputee is in a wheelchair. To determine how mobile an amputee is with their prosthesis, an outcome measure must test some weight-bearing activities: standing, walking or running. Therefore, deciding which activities to measure is central to demonstrating the effectiveness of a prosthesis on the pwLLA's ability to mobilise, i.e. the construct being measured by the outcome measure must be a construct of interest to the researcher or clinician.

There does not appear to be a clear consensus on which outcome measure(s) to use, within clinical groups working in prosthetic rehabilitation; this is despite a wide variety of outcome measures purporting to measure aspects of physical function to choose from. The choice requires an understanding, not only of the construct the instrument is designed to measure, but also its measurement properties i.e. its reliability, validity and responsiveness. A full understanding of these measurement properties, also known as psychometric properties, will help clinicians and researchers choose the most appropriate outcome measure. For example, when assessing the differences between one prosthetic component and another it is critical to know whether an outcome measure is able to detect changes accurately in the performance of a pwLLA. The use of valid and reliable outcome measures, that

are responsive to change, can also objectively record the functional ability of a pwLLA at each stage of their rehabilitation.

Previous reviews of the literature have highlighted a lack of evidence available on the psychometric properties of outcome measures used with pwLLAs (Condie et al 2006, Deathe et al 2009, Hebert et al 2009). There are also concerns that current prescription guidelines (e.g. Medicare K levels) may be too rigid when used in the standardisation of the provision of prosthetic components by not allowing lower functioning amputee's access to higher functioning components. There have been calls for more objective measures of ability in the prescription process (Kannenberg et al. 2014), where the use of more robust outcome measures in the assessment phase, could lead to improved performance and safety benefits for the patient.

The research topic of this PhD was prompted by an interest in guiding clinicians and researchers on the appropriate use of outcome measures during prosthetic rehabilitation. In particular, those outcome measures that measure physical performance of pwLLAs. The premise being that by providing robust evidence of the outcome measures that are used, clinicians and researchers will be able to effectively measure the impact of prosthetic componentry. In the context of rising healthcare costs this will assist in getting value for money for all pwLLAs.

1.2 Topics and key terms

1.2.1 Epidemiology and classification of lower limb amputations

Lower limb amputees are often classified as having had a major or minor amputation. This relates to the location of the surgery, and the most common major amputation levels of the lower limb are; trans-tibial (TT), trans-femoral (TF), through-ankle or Symes amputation, through-knee and through-hip or hip disarticulation. Other amputations below the ankle, either through the mid or fore-foot or amputation of any or all of the toes, are classified as minor. In many rehabilitation studies involving pwLLAs, these minor level amputations are usually reported together as a group because of the small numbers recorded.

Further discussion of the aetiology and care-pathway of a pwLLA from surgery to the provision of their prostheses, will be presented in chapter 2.

1.2.2 Outcome measurement

Outcome measures have been described in various ways as: instruments, scales, tools, indices etc. and have been designed to measure a huge variety of psychological, physiological and physical characteristics of human behaviour, attitudes and activities. These characteristics can be measured in order to classify, predict, and motivate individuals, as well as attest to the success (or otherwise) of healthcare interventions. Within the field of prosthetic rehabilitation, amputees, clinicians, researchers, managers and policy makers all use outcome measures for a diversity of reasons.

The way information (data) is collected from the patient varies depending on the type of outcome measure. Outcome measures that ask the individual to report on their own capacity or capability or how they feel about their accomplishments are often referred to as self-reported or patient-reported outcome measures (PROMs). Performance-based measures are more objective as the activities are recorded by an observer. These are the two main types of outcome measure that will be discussed in this thesis.

The topic of outcome measurement will be further discussed in detail in chapter 2, with evidence, demonstrating which outcome measures are currently used within the clinical practice setting of prosthetic rehabilitation, presented in chapter 3.

1.2.3 Psychometrics vs Clinimetrics

The term psychometrics in relation to outcome measures is used to describe the measurement properties (i.e. internal consistency, reproducibility or reliability, validity and responsiveness) of the outcome measures and was developed in the field of psychology (Streiner 2003). Clinimetrics is a term that is also used to describe measurement properties, however its base is within the clinical or health science fields and is employed in relation to instruments utilised in those areas. It has been argued that clinimetrics widens the range of information with the inclusion of “soft data” (Fava et al. 2012). When considering the role of clinical assessment in

the 21st Century, soft data, including judgements on impairment and well-being, are a necessary addition to the hard data obtained from laboratory results. The term clinimetrics has only begun to be used relatively recently, whereas psychometrics has been in use far longer. There has been a call to stop using the term clinimetrics, to clear up any confusion that may arise (Streiner 2003), as both terms are regularly used synonymously throughout the literature. Galea in 2005 published this explanation "the purpose of clinimetrics is to alert clinicians to the psychometric properties of instruments that have clinical utility in current physiotherapy practice" (Galea 2005). In an effort to minimise confusion, the term psychometrics will be used throughout this thesis.

1.2.3.1 Psychometric properties

A variety of terms can be used when describing the psychometric properties of an outcome measures. A Delphi study undertaken by a group from the COSMIN initiative (COnsensus-based Standards for the selection of health Measurement INstruments) aimed to reach a consensus on a) the relevant psychometric (measurement) properties for health-related patient-reported outcomes (HR-PROs) and b) the terminology and definitions of these measurement properties (Mokkink et al 2010a). The authors proposed a taxonomy of terms, with the measurement properties grouped within three main domains: Reliability, Validity and Responsiveness. There appeared to be most debate about reliability, but the terminology agreed on were; internal consistency, reliability and measurement error all combined under the domain of reliability. The other two main domains were: validity, which includes the measurement properties; content validity, face validity, construct validity, structural validity and criterion validity and; responsiveness.

While the outcome measures discussed in this thesis will also include observed measures of performance, and not only HR-PROs, the terminology and definitions published by Mokkink et al in 2010 will be used throughout. Further discussion of the psychometric properties that are pertinent to this thesis will be presented in Chapter 2. Study design requirements and preferred statistical methods for evaluating HR-PROs was an additional aim of the Delphi study undertaken by Mokkink et al (2010) and will also be discussed in more detail in relation to the systematic review in Chapter 4.

1.2.4 Conceptual framework

An outcome measurement framework will be considered throughout this thesis together with the concept that treatments / interventions change performance.

Within the International Classification of Functioning Disability and Health (ICF) framework if an individual has a problem executing a task or action they are classified as having a limitation at the person level. If they are limited by an inability to perform a task to participate or be involved in personal, social or work activities they are deemed to have limitations at a societal level. Prosthetic rehabilitation as an intervention, will impact on activity limitations for a pwLLA. However the bio-psychosocial model of disability, that is the basis for ICF, interlinks activity at an individual level with impairments at the body level and also limitations at a societal level (Ustun et al. 2003).

Further observations of the ICF classification and its uses with regard to the pwLLA population will be explored more fully in Chapter 2. Examining the outcome measures used in prosthetic rehabilitation and the evidence of their psychometric properties, will also be more fully explored and addressed in later chapters (3, 4 and 5).

1.3 Structure of thesis

The structure of the thesis is outlined in Figure 1.1. The thesis includes a background chapter followed by four chapters, each of which will present the methodology, results, discussion and conclusion separately for the four investigative and empirical studies which comprise the main body of the thesis. The final discussion chapter brings together the conclusions from each study and presents their implications in the context of the prosthetic rehabilitation and research; as well as in the context of the wider debate on outcome measurement.

An overview of the four studies within the context of the research aims and specific questions posed within this thesis will be presented at the end of chapter 2.

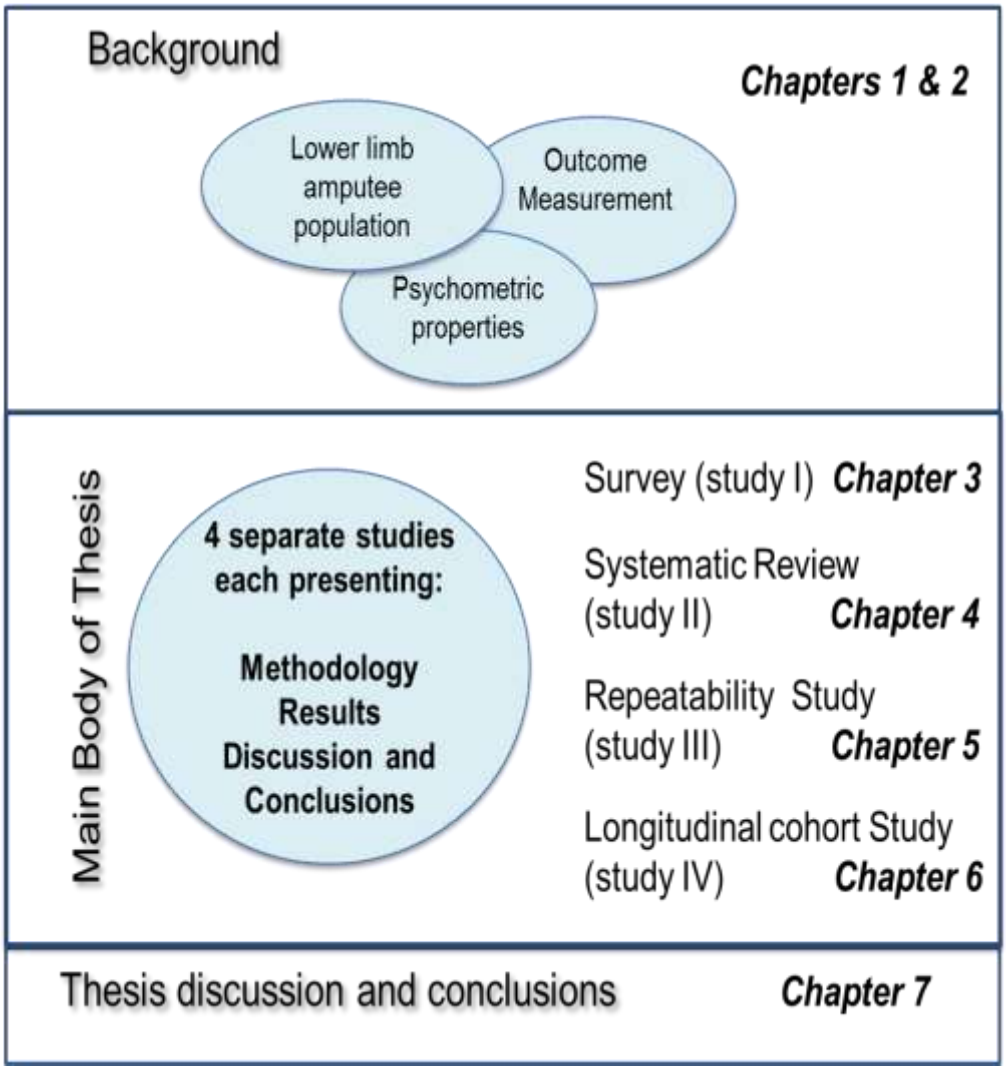


Figure 1-1 Structure of thesis

2 Chapter 2 Background

2.1 Purpose of chapter

The purpose of this chapter is to summarise the evidence on which the research questions of this thesis are based. The essential features of the literature, as it relates to this thesis, will be presented. Key gaps in the knowledge base identified will be discussed to support the rationale for the thesis.

2.2 Lower limb amputee population

This section will introduce the population of pwLLA and explore the pathway of care from the decision to amputate through to fitting of a prosthetic limb (limb-fitting).

2.2.1 Epidemiology and aetiology

Lower limb amputation may be undertaken for a variety of reasons, the most prevalent of which is peripheral arterial disease (PAD). Atherosclerosis of the peripheral arterial vessels restricts the supply of blood to the extremities and, if not revascularised, ischaemia and necrosis of the tissues will eventually lead to amputation. End-stage renal disease, poor functional status and diabetes increase the need for amputation in patients with PAD (Dillingham et al. 2005). Those patients with diabetes, in particular, have an increased risk for amputation compared to those who do not have the disease (Jeffcoate et al. 2012). In addition, not only are diabetics likely to experience their initial amputation at a younger age (Davie-Smith et al. 2016) but they are also more likely to progress to higher-level amputations (Holman et al. 2012). Further complications are also to be expected as they are likely to be more severely disabled and die at a younger age compared to amputees without diabetes (Dillingham et al. 2005). An epidemiological study of lower limb amputations in England confirmed these findings with results showing that 43.1% of patients undergoing an amputation had a primary diagnosis of cardiovascular disease, 12.3% had endocrine, nutritional or metabolic disease, which includes diabetes, 13.9% had injury or trauma, 2.2% neoplasm and 6.9% were unclassified (Moxey et al. 2011, Moxey et al. 2010). The same study showed that 39.4% of patients undergoing major lower limb amputations had diabetes and the percentage rose to 50% in those undergoing minor amputation (below the ankle

joint). Similar overall prevalence data was found in a retrospective audit of hospital data (Ahmad et al. 2014). The prevalence of major lower limb amputations in men was seen to be twice that of women in the Ahmad study and also in a retrospective cohort study of non-traumatic patients undergoing amputation in Scotland (Davie-Smith et al. 2016).

The Scottish Physiotherapy Amputee Research Group (SPARG) collect data on all amputations carried out in Scotland. The latest data, published in March 2016, giving figures for the year 2013, showed there were 848 amputations (on 809 patients) during that year (Scott et al. 2016). The demographic data reported, is broadly similar to the breakdown of the English data, with an average age at amputation of 67 years (median 70) and approximately two thirds (68%) male. However, with 85% of all amputations due to PAD and/or diabetes, this may reflect the higher prevalence of cardiovascular disease in the Scottish population (Morris et al. 2001, Leyland 2005). The percentage of TT amputations was 56% and 40% were at the TF level.

It is not surprising then that together, PAD and diabetes account for the majority of cases referred for lower limb prosthetic assessment in the UK (57%) with infections (10%) and trauma (9%) the next most frequent causes (United National Institute for Prosthetics,& Orthotics Development 2013). The most recent report from the United National Institute for Prosthetics & Orthotics Development (UNIPOD) managed by the University of Salford showed that there were nearly 6,000 (5906) new amputee referrals for prosthetic treatment in 2011-2012 (UNIPOD 2013). Lower limb amputations contributed to 91% of these referrals, with 53% aged 64 years or older and just over one quarter (27%) aged over 74. The highest number of referrals were for patients with a TT amputation (56%) and 48% of the referrals were for TF amputations. These figures are similar to previous year's referral data. Although the average age of those referred for to the limb-fitting centres is not presented, age ranges show 53% of all referrals are over the age of 64 and 27% are over the age of 74.

While there has been a large amount of publicity surrounding the military personnel who have undergone amputations as a result of recent conflicts their numbers are relatively small. Data from the UK Ministry of Defence showed a peak of 82

amputations in 2010 - 2011 but the annual figures are now falling, with 55 reported in 2011-12 and 40 in 2012-13 and no more than 10 each year since (Defence 2015). In contrast to the civilian population, the military amputee is much younger and with fewer chronic co-morbidities. However, the demands that these younger and more active amputees put on their prostheses have helped to promote advances in prosthetic manufacturing and rehabilitation for all pwLLAs (Laferrier and Gailey 2010).

2.2.2 From surgery to limb fitting

2.2.2.1 Pre-operative assessments and decisions on amputation level

When an amputation is undertaken the aim is to preserve as much function as possible. This is done by utilising the most viable tissue and saving as many joints as possible. However, there is a fine balance to be reached on the amount of viable tissue with a healthy blood supply that can be retained, allowing rehabilitation without risking further surgery or the spread of infection in vulnerable tissue. Which surgery is undertaken, and when, may have a considerable effect on the resulting physical capability of a pwLLA. Evidence from a recent systematic review indicated that skin problems affecting tissue viability appear to be linked to poor tolerability of the prosthesis (Butler et al. 2014). The authors noted a lack of robust evidence on the long-term effects of poor tissue viability and recommended further work is undertaken. Identification of clearer links between tissue viability and other key factors may help inform help clinical management strategies and care pathways for pwLLAs.

An amputation at the TT level obviously preserves the knee and hip joints, while a TF amputation preserves only the hip joint. The resultant energy requirements to propel a prosthesis are increased at the TF level (Genin et al. 2008). With the majority of amputees over 60 years of age, this will therefore have an impact on their level of activity with a prosthesis. Amputations that bisect a joint, cutting through only the soft tissues around the joint, are quicker to perform, e.g. at the knee joint the femoral condyles are left intact leaving a longer lever which helps provide more stability in sitting (Morse et al. 2008). This type of surgery may be considered if the patient is an anaesthetic risk and not necessarily a candidate for

rehabilitation with a prosthesis as this particular level is not a popular choice with prosthetists as there are challenges when fitting a prosthetic knee joint.

Pre-surgery consultations with the prosthetic rehabilitation multi-disciplinary team (MDT) can speed up the post-operative recovery period by assessing the patient early, and in some cases starting the rehabilitation programme before surgery. Establishing the physical condition of a pwLLA pre-operatively will help in this recovery process, as increasing age at time of amputation, a high number of co-morbidities and poor physical function have been shown to negatively impact on the prosthetic outcome (Munin et al. 2001, Schoppen et al. 2003, Hamamura et al. 2009). Knowledge of these, and other factors, help predict whether the pwLLA will be a successful prosthetic limb-user; as well as the likely activity level they will reach, e.g. indoor / outdoor walking, or more active pursuits (Taylor et al. 2005).

2.2.2.2 Early post-amputation rehabilitation

Many factors, both physical and psychosocial will influence the outcome of the rehabilitation (Schaffalitzky et al. 2012, Desmond et al. 2008, Gallagher et al. 2011), and not all pwLLAs will benefit or be able to use a prosthetic limb. From the SPARG data collected in 2013 it was seen that 322 (41%) pwLLAs progressed to prosthetic limb-fitting following their amputation and 22 (3%) subsequently abandoned their prosthesis before discharge. Forty-six percent of pwLLAs (365) were not limb-fitted and the remainder either died or were lost to follow-up (Scott et al. 2016).

Before a decision is made whether a prosthetic limb will be prescribed, the amputee is assessed by the MDT for their suitability to use a prosthesis. The use of an early walking aid (EWA) will help assist with this decision. There are several types of EWA available, but the principle is to get the LLA walking on a temporary leg, as soon as possible after the surgery. This will guide the MDT, and the amputee, on whether they have the physical capabilities to use a prosthesis for mobilising. Specially designed outcome measures have been developed to assess the LLA for the purpose of predicting their physical potential as a successful limb-user or not, for example; the Trans-femoral Fitting Predictor (TFP) tool (Condie et al. 2011), and the Amputee Mobility Predictor (AMP) tool (Gailey et al. 2002, Raya et al. 2013). It is often decided, that given the average age and range of co-morbidities of the majority of pwLLAs, many will have a better quality of life using a wheelchair, especially if

they have had a TF amputation due to the increased energy expenditure required to mobilise at that level (Genin et al. 2008, Goktepe et al. 2010).

2.2.2.3 Limb-fitting and prosthetic rehabilitation

If an amputee is deemed to be a suitable candidate for a prosthesis, the MDT must then consider which prosthetic components will suit the predicted activity levels for the amputee. Whether the pwLLA is receiving their first (primary) prosthetic limb, or they are more established using their prosthesis and are looking for a replacement, or a second limb is required for other activities, e.g. swimming or running, the same physical considerations are taken into account. These include; what the overall current clinical/medical condition of the pwLLA is, and the potential for that to change; how active they are and what potential improvement the prosthetic limb will have on their lives. This last aspect is often the most important to the amputee but is the hardest to predict and measure.

2.2.3 Classification of functional ability and prosthetic prescription

It is fundamental to the prescription process that both the individual's activity and capability levels are assessed when choosing the correct prosthetic component for that individual. The choice of prosthetic component is often guided by the level of the pwLLA's functional ability and the use of a classification system assists in this process. In 2001, the US Health Care Financing Administration published a 5-level functional classification system using code modifiers (K0, K1, K2, K3, K4), known as the Medicare K-level classification system (see Table 2.1). Each "K-level" relates to the functional abilities of the pwLLA, and the lower the activity potential of the amputee, the lower is his/her assigned K-level and vice versa.

Table 2.1 K-Level classification (adapted from AAO&P website)

K-Level	Description
K0	This patient does not have the ability or potential to ambulate or transfer safely with or without assistance and a prosthesis does not enhance their quality of life or mobility.
K1	This patient has the ability or potential to use a prosthesis for transfers or ambulation on level surfaces at fixed cadence - a typical limited or unlimited household ambulator.
K2	This patient has the ability or potential for ambulation with the ability to traverse low-level environmental barriers such as curbs, stairs, or uneven surfaces - a typical community ambulator.
K3	The patient has the ability or potential for ambulation with variable cadence - a typical community ambulator with the ability to traverse most environmental barriers and may have vocational, therapeutic, or exercise activity that demands prosthetic use beyond simple locomotion
K4	The patient has the ability or potential for prosthetic ambulation that exceeds basic ambulation skills, exhibiting high impact, stress, or energy levels - typical of the prosthetic demands of the child, active adult, or athlete

These levels are arrived at by questioning patients as to their current functional ability as well as taking into account their concurrent medical condition and trying to predict what level they may achieve. There does not appear to be an agreed method of gathering information on activity levels nor which specific data should be recorded (van Twillert et al. 2013). However, this classification system is widely used to guide the reimbursement of healthcare payments across the United States. It is also used by most prosthetic companies to classify the use of different prosthetic components, such as feet and knee joints. It should be acknowledged that there appears to be an absence of any formal assessment of psycho-social factors when considering the prosthetic prescription for a pwLLA. While this is worthy of further discussion, there will be no in-depth handling of that topic here, as the main focus of the thesis is the examination of outcome measures of physical function used during prosthetic rehabilitation.

Regulating the choice of prosthetic components on the basis of functional ability represents consideration of the cost-effectiveness of prescribing such components

to less active users. However, evidence is emerging that some of the newer, more expensive, components may be beneficial for the frailer amputee rather than the more active pwLLA for whom they are usually prescribed. For example, MCPKs were introduced primarily for the more active amputee at the high K3 and above level. However, some benefits, such as “stumble control” have been observed with less active amputees at the K2 level (Sawers and Hafner 2013, Kannenberg et al. 2014). The results of both reviews however, are to be viewed with some caution with regard to the outcome measures used. The evidence provided with regard to activity levels in the Sawers & Hafner review came from eight studies. The authors recognised the use of appropriate outcome measures was essential to a good quality study and had therefore noted whether valid and reliable outcome measures were used. However, they concluded that in only two of the eight studies, the use of (what they considered) a reliable outcome measure was identified (Sawers and Hafner 2013). In the review by Kannenberg et al published in 2014, the authors identified that current classification systems have an element of subjectivity that calls for speculation of the amputee’s functional potential, by the assessor, and noted this as a limitation of the studies included. They have called for a validated and “unambiguously quantifiable” classification system that could objectively assign functional levels and be used with performance-based tests to identify clinically meaningful improvements.

Fortington et al (2012) also identified that a mobility outcome measure should be valid and reliable as well as being responsive to change when they looked at whether a prognosis for mobility (prosthetic or non-prosthetic) could be made in elderly pwLLAs. However, their findings were inconclusive because of “incomplete reporting of study populations and poor reporting of the reliability of the mobility measures used”. They recommended further research into the mobility outcomes of this elderly population (Fortington et al. 2012).

The use of valid and reliable outcome measures that are responsive to change and can objectively record the functional ability of the pwLLA at each stage of their rehabilitation, would help standardise the provision of prosthetic components. This standardisation could play an important role in the clinical decision making process of prosthetic prescription. However, it would seem that there is a gap in the evidence provided for such outcome measures.

2.2.4 Outcome measures used with pwLLAs

While the focus of this thesis is on physical function, in the context of prosthetic rehabilitation, many outcome measures are employed to assess different aspects of the individual patient's progress as well as the effect of therapy programmes or changes in prosthetic components. However, research studies investigating these effects will often report results only at the group level. As a result clinicians will not have information on how many outcome measures may be used at the individual level in clinical practice.

2.2.4.1 Amputee-specific outcome measures

Despite there being a large number of general outcome measures available, many specific amputee outcome measures have been developed. These outcome measures have been specifically designed for use with pwLLAs for a variety of reasons. See table 2.2 for those that have been developed and subsequently described in published studies. While others were developed with different populations and have been subsequently validated for use with amputees, see Table 2.3.

Table 2.2 Outcome measures specifically developed for pwLLAs

Outcome Measure	Type	Main Constructs being Measured
Amputation Related Body Image Scale	PROM	Body Image
Amputee Activity Score (AAS)	PROM	Functional Activities
Amputee Mobility Predictor (AMP)	Observed	Functional Activities
Functional Measure for Amputees (FMA)	PROM	Functional Activities
Houghton Scale	PROM	Prosthetic Use
Locomotor Capabilities Index (LCI) or Modified LCI-5	PROM	Functional Mobility
Orthotics and Prosthetics National Outcomes Tool (OPOT)	Mixed	Health-related Quality of Life (HR-QoL), Functional Mobility
Prosthesis Evaluation Questionnaire (PEQ)	PROM	Prosthetic-related Quality of Life (PR-QoL)
Prosthetic Profile of the Amputee (PPA)	PROM	PR-QoL, Functional Activities
Prosthetic Observational Gait Score (POGS)	Observed	Gait

Russek's Code	Observed	Functional Ability
Special Interest Group in Amputee Medicine (SIGAM) Mobility Grades	PROM	Functional Mobility
Socket Comfort Score (SCS)	PROM	Pain / Discomfort
Trinity Amputation and Prosthesis Experience Scales (TAPES)	PROM	HR-QoL

Some of these have been developed to measure a specific feature or characteristic to the amputee e.g. comfort of the socket while others have been developed because it was felt that outcome measure that did exist, did not address particular issues well enough for the needs of pwLLAs.

2.2.4.2 General outcome measures

Many general outcome measures that were developed, initially for use in other patient populations, have subsequently been validated for use with amputees. See table 2.3 for a selection of general outcome measures which have been have been reported in the literature as being used in pwLLA.

Table 2.3 General outcome measures used with pwLLAs

Outcome Measure	Type	Main Constructs being Measured
Activities-specific Balance Confidence (ABC) Scale-UK	PROM	Balance, Functional Mobility
Barthel Index	Mixed	Activities of Daily Living (ADL)
Berg Balance Score (BBS)	Observed	Dynamic and static balance
Body Image Questionnaire	PROM	Body Image
Canadian Occupational Performance Measure (COPM)	Mixed	Functional Activities
Frenchay Activities Index (FAI)	PROM	Functional Mobility, ADL, Occupational performance
Functional Independence Measure (FIM)	Observed	Functional Activities
Patient Generated Index	PROM	ADL, General Health, Motor Activities
Perceived Social Stigma Scale	PROM	HR-QoL
Rivermead Mobility Index (RMI)	Mixed	Functional Mobility, Balance
Short Form 12 or 36 Health Survey (SF-12 or	PROM	HR-QoL

36)		
Sickness Impact Profile (SIP)	PROM	HR-QoL, ADL
Timed “Up and Go” Test (TUG)	Observed	Mobility and Balance
Timed Walk Test (TWT) - 2min (2MWT) or 6min (6MWT)	Observed	Mobility Endurance

An extensive number of outcome measures have been used with pwLLAs for a wide variety of reasons, and it has been stated on numerous occasions that there is no consensus about which outcome measures should be used (Rommers et al. 2001, Condie et al. 2006, Gremeaux et al. 2012). There is also no consensus on any that may be considered gold standard measures within this population.

The Core Outcome Measures in Effectiveness Trials (COMET) initiative promotes the importance of agreeing a standard set of outcomes that represent the minimum that should be reported in all clinical trials (Williamson et al. 2011, Prinsen et al. 2014). Reaching a consensus on which data should be collected, allows the results to be compared and combined, from research studies and clinical audit, as appropriate.

All physiotherapists working in prosthetic rehabilitation in Scotland regularly collect outcome data on the amputees coming through their service and these data informs the SPARG annual report (Scott et al. 2016). The modified Locomotor Capability Index (LCI-5) is collected routinely at the end of the initial period of prosthetic rehabilitation. There is yet to be multi-disciplinary agreement across the care team on which outcomes should be collected. However, both the Physiotherapy and Prosthetic Professional bodies in the UK have independently brought out a list of recommended outcome measures to be used in prosthetic rehabilitation, (Cole et al. 2014, Young et al. 2015).

The International Society for Prosthetic and Orthotics (ISPO), which is a multi-disciplinary organisation, has published recommendations, for defining participants in prosthetics research (Lemaire 2013). They outline measures that can be included such as: Manual Muscle Testing (grades 0-5), walking speed (m/s), Steps taken (step/min), Socket interfacial pressure (kPa) and “other” kinematic and kinetic gait parameters, though which are not specified.

The continued lack of a consensus of practice, with regard to the collection of outcome measure data has led to the first research question in this thesis:

Research Q1
What outcome measures are used regularly in clinical practice during prosthetic rehabilitation?

This question will be addressed in Chapter 3.

2.3 Outcome measurement

Psychometric properties of outcome measures will be explored in this section of the chapter. Understanding what psychometric properties are will help clinicians and researchers interpret results obtained from outcome measures more fully. An exploration of the philosophy and literature regarding aspects of measuring outcomes in general terms will be presented first. Then finally, the ICF framework will be described, outlining its role in prosthetic rehabilitation.

2.3.1 Conceptual framework and definitions of outcome measurement

The Donabedian Theoretical Model first described in 1966 (Hoenig et al. 2010) provides a framework for systematic investigations in healthcare. The model is made up of three parts; firstly, the Structure of Care which includes the setting where the healthcare is being delivered, the material resources and the organization structure; secondly, the Process of Care i.e. what is actually being done during the giving and receiving of care, and; finally, the Outcomes of Care which describe the effect of the care on the health status of not only an individual but also the wider population. Instruments that capture those effects are referred to as outcome measures.

Outcome measures in healthcare often identify reaching certain milestones or achieving certain scores. These milestones and scores may then be quoted when the success, or otherwise, of a service, a treatment, a programme of therapy or an intervention, is discussed. During the course of an illness, clinicians, and their patients, often want to keep track of any changes that occur. Implementation of

therapy, or any other kind of intervention, may change the condition and again clinicians as well as their patients will want to track the effects. The use of outcome measures will help do this, if they are used in the right context for the right reason. They also must be “fit for purpose”. In a review of health measurement, Cano and Hobart in 2011 argued that the evidence for a large number of outcome measures remained unclear on that point (Cano et al. 2011).

During rehabilitation outcome measures may be utilised for several reasons (Stokes 2010). Some are used to classify the patient’s current condition, knowledge of which can help predict the success or failure of a procedure, treatment programme or intervention. For example, lung function tests prior to an exercise programme for a patient with a respiratory condition, will not only categorise the level of function the patient has, but will also help predict their level of participation and how quickly the programme may need to be adjusted. Other outcome measures are used to monitor progress during the rehabilitation period, e.g. balance tests undertaken by a frail elderly patient, before, during and after during a therapy programme, will help the clinician target the use of specific exercises, as well as identify if the therapy is helping to improve the patient’s balance. Finally, outcome measures can also be used to comment on or classify the success, or otherwise, of an episode of care for a group of patients.

It is important to understand what the purpose or target of any outcome measure is, so that their results may be interpreted correctly (Roach 2006). For example, with a focus on outcome measures of physical function in this thesis, physical ability will therefore be one of the targets or constructs of interest for any outcome measure being studied.

2.3.2 Psychometric properties

The study of psychometrics considers the different measurement properties of outcome measures, how they are defined and analysed. The correct interpretation of any results obtained from an outcome measure, relies on that outcome measure being valid, reliable and responsive to changes in the performance capability of the population of interest, i.e. pwLLAs. This section will outline the different psychometric terms used and various statistical analysis methods that are employed to establish the strength of each measurement property.

2.3.2.1 Taxonomy of terms

Reliability and validity are arguably, the most familiar terms used to define what characteristics a "good" outcome measure should have. Responsiveness, or sensitivity to change, are also terms that are becoming more familiar, especially when considering if an outcome measure is effective in detecting changes in the patient's condition or their performance. The COSMIN initiative has sought to gain agreement on the definitions of terms used through a Delphi study (Mokkink et al. 2010a). The group have also proposed criteria, on the reporting of these measurement properties (Mokkink et al. 2010b, Mokkink et al. 2010c). The purpose is to enhance the quality of the studies being published and therefore improve the level of the evidence for clinicians and researchers. The quality criteria will discussed in more detail later in this chapter. Table 2.4 lists the definitions proposed by the COSMIN group.

These terms will be used throughout this thesis. Further clarification on each of the major domains and the use of different statistical analysis methods to establish the psychometric properties of outcome measures, will be presented in more detail in the following sections. Particular attention will be addressed to the topic of establishing reliability, measurement error and responsiveness.

Table 2.4 Domains and definitions of measurement properties (adapted from Mokkink et al 2010b)

Domain	Measurement property (*specific aspect*)	Definition
Reliability		The degree to which the measurement is free from measurement error : The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions, e.g. using different sets of items from the same HR-PROs (internal consistency), over time (test re-test) by different persons on the same occasion (inter-rater) or by the same persons (i.e., raters or responders) on different occasions (intra-rater)
	Reliability	The proportion of the total variance in the measurements which is because of “true” differences among patients
	Internal Consistency	The degree of the interrelatedness among the items.
	Measurement error	The systematic and random error of a patient’s score that is not attributed to true changes in the construct to be measured.
Validity		The degree to which an instrument measures the construct(s) it purports to measure.
	Content (*Face*)	The degree to which the content of an instrument is an adequate reflection of the construct to be measured. (*degree to which (the items of) an instrument indeed looks as though they are an adequate reflection of the construct to be measured*)
	Construct (*Structural*)	The degree to which the scores of an instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the instrument validly measures the construct to be measured. (*degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured*)
	Criterion	The degree to which the scores of an instrument are an adequate reflection of a ‘‘gold standard’’
Responsiveness		The ability of an instrument to detect change over time in the construct to be measured

2.3.2.2 Reliability

The overall definition for the reliability domain is; the degree to which the measurement is free from measurement error and the extent to which scores for patients that have not changed are the same for repeated measurement under several conditions. The three sub-divisions in this domain address slightly different aspects of reliability. **Reliability** determines the extent to which scores (for patients who have not changed) are the same for repeated measurement under the same conditions. The consistency and agreement of the scores are considered in this measurement property. **Internal consistency** considers how related items are within the same measure. It is used to comment on the inter-correlation of items in a scale or sub-scale and to what extent they are measuring the same construct or concept. Internal consistency (IC) parameters have also been reported as evidence of the reliability of a test, however IC is based on a single occurrence and the variances that occur from day to day or between different raters are not taken into account (Tavakol and Dennick 2011). Finally **measurement error** considers the "noise" of the instrument or outcome measure. This "noise" consists of systematic and random errors that are part of a patient's score, not attributed to any true changes, in the construct to be measured. Reliability parameters are highly dependent on the variation in the population sample. Therefore they are a characteristic of the outcome measure in a certain population (Weir 2005).

Reliability and measurement error will now be considered in more detail given the focus of this thesis.

2.3.2.2.1 Reliability

This is the reproducibility of results when a test is repeated more than once on subjects over time (test-retest) or; implemented by different persons on the same occasion (inter-rater) or; by the same persons (i.e. raters or responders) on different occasions (intra-rater).

The consistency of the results is considered the degree to which the results do not change when the measure is repeated, whereas agreement refers to the closeness of the results obtained to the "true" result and is free from measurement error. The word ''true'' must be seen in the context of the Classic Test Theory (CTT), which

states that; any observation is composed of two components the true score and error associated with the observation. ‘‘True’’ is the average score that would be obtained if the scale were given an infinite number of times (Streiner et al. 2014).

2.3.2.2.2 Considering consistency

If a research question asks whether an outcome measure is able to distinguish between subjects and / or raters then reliability parameters should be investigated, i.e. how consistent are the results for different raters or different subjects. Intra-class correlations (ICC), based on analysis of variance (ANOVA) are the most frequently used method to report reliability of a measure in terms of its consistency. Correlations give an indication of the relationship between two sets of results. The ICC value obtained, between -1 and + 1, will demonstrate not only the strength of the relationship but also the direction. Zero indicates no relationship and the nearer to 1 the stronger the relationship in a positive (+) or negative (-) direction.

The precise ICC model used is very often not quoted. Shrout and Fleiss (1979) outline some guidance for which model to choose (Shrout and Fleiss 1979). They suggest considering the following questions before deciding which model to choose: Firstly, does the statistical model require one-way or two-way analysis of variance? Secondly, are differences between the mean scores or ratings from several raters relevant to the reliability being studied? Finally, is the analysis a unit of analysis an individual rating or the mean of several ratings? The first two questions relate to the statistical model being considered. The second two questions are related to the possible use of the results.

With several models to choose from, e.g. 1,1, 2,1, 1,k, 2,k, where k reflects the number of raters or tests included in the analysis, it is important to choose the correct one. If there are several raters who each rate some of the total number of subjects but they don’t all rate all of the subjects, then a one-way ANOVA should be used, i.e. models 1,1 or 1,k . This does not take into account the variance attributed to the raters, because they haven’t rated all the subjects and therefore it is not accounted for across all the results. However, if each of the subjects are rated by all raters (this may be a number of raters taken at random from a larger population of raters or where there is only one rater rating all the subjects) then the choice should be a two-way ANOVA, i.e. 2,1 or 2,k. This will take into account the additional variance of multiple raters across the results. If a one-way ANOVA is used instead

in this case then the resulting correlation would be an under-estimation of the relationship (Streiner et al. 2014).

If the effect of the rater(s) is considered random and results are to be generalised to other raters, then models 2,1 or 2,k should be used. If on the other hand the effects are not considered random then models 2,2 or 2,k should be chosen. This implies that the results will only be concerned with a single rater or defined group of raters and the results will not be generalizable to the wider population of raters. If model 2,1 (the reliability unit of analysis is a single rating and not the mean of several), is used in preference to 2,2 then it is likely that this will result in a lower correlation value (Shrout and Fleiss 1979).

Two other methods of analysis are sometimes used to report consistency; Pearson correlation and Kappa coefficient (Streiner et al. 2014).

Pearson correlation is a regression analysis of pairs of results, commenting on the relationship between the two sets of results. There is a suggestion that coefficient values obtained from this method are likely to be higher than true reliability. As ICC account for the variances the ICC values will be closer to “true” reliability. Another benefit to using ICC over Pearson is that the calculations can cope with multiple raters or observers.

Kappa coefficient is the most appropriate when only two levels are expected, e.g. when something is either present or absent in a test result. Weighted Kappa can be used when a scale of results is expected, e.g. a 6 point rating scale of muscle strength. This will take into account partial agreement where results differ by one or two points on the scale and so in this instance weighted Kappa is looking at disagreements. Weighting schemes have been published and are recommended over using individual schemes to allow comparisons to be made with other studies. However there are similarities between the weighted Kappa and the ICC and the overall the decision over which to use may come down to how easy each calculation is to perform and preference (de Vet et al. 2011).

2.3.2.2.3 Considering agreement

Rankin and Stokes (1998) discuss limits of agreement and how they should be used with ICCs. They suggest neither should be used alone as they present two different, but related, aspects of reliability (Rankin and Stokes 1998).

There are several agreement parameters that can be calculated for an outcome measure, see table 2.5.

Table 2.5 Description of agreement parameter indices (Streiner et al. 2014, Bland and Altman 1986)

Agreement Parameter Indices	**Description**
Standard Error of Measurement (SEM)	Represents the variance between subjects. Depending on the method of calculation will take in to account systematic error.
Limits of Agreement (LoA)	Presented in graphic form where outliers will be obvious Separates the effect of random error from the main effect of the measure.

Reliability coefficients involve error variance, also known as standard error of measurement (SEM) and variance between subjects, often presented as standard deviation (SD). By reworking the calculations, the SEM of a measure can be presented. Standard error of measurement is expressed in the unit of the outcome measure and allows a confidence interval to be drawn around a score (Stratford 2004). Whether this is the observed or "true" score, i.e. with systematic error accounted for, depends on the method of calculation.

Bland and Altman (1986) present a method of examining agreement in an outcome measure that does not take account of the true variability of observations. The Bland Altman method uses similar calculated values to those used in ICCs and this absolute measure of agreement separates the main effect of the measure from random error. The graph or plot that is produced is integral to the reporting of the method. By plotting the mean difference of each participant's data, against the difference in their scores on the same two occasions, any systematic differences or increase in error that may be related to the value of the measurement are seen graphically. Any obvious outliers can also be seen in the graph (Bland and Altman 1986). The limits of agreement (LoA) that are represented on the graph are also

approximately twice the measurement error values that are reported for the same population.

Agreement is more a characteristic of the instrument itself and is preferable in all situations where the instrument will be used for evaluative purposes. It is the degree to which scores or ratings are identical irrespective of who performs or scores the test. However, if the question concerns a subject's health status and measuring any changes in that, then agreement parameters should be presented, as the main concern is ensuring that the result is accurate. Arguably, both consistency and agreement should be presented to give full account of the measurement properties for the outcome measures and reflect both the consistency and accuracy of the instrument (de Vet et al. 2006, Kottner et al. 2011).

2.3.2.2.4 Measurement error

In a health care setting, we are often concerned with measuring change in health status. Any systematic and random errors that are not attributed to true changes in the construct to be measured must be taken into account. The threshold of error is inherent to the population and condition being studied and for well-established tests this threshold is well known, e.g. weighing scales will be accurate to 1-2kg which is deemed acceptable for adults where daily weights may fluctuate to this extent, however this value would not be appropriate for infants or adults who were on dialysis. The clinical interpretation of SEM can be expressed as the minimal detectable change (MDC) value or smallest detectable change (SDC) (Haley and Fragala-Pinkham 2006).

Reliability and agreement parameters are population specific and to establish error values for new tests, multiple data points must be recorded for many patients of different ages who have demonstrated different levels of change and with different conditions if the test is to be used across populations, conditions and age ranges (Kottner et al. 2011).

2.3.2.3 Validity

In the past, validity was only seen as testing the psychometric properties of a given scale or measurement tool. However, there is now a shift towards focusing on the

subjects who perform or complete the scales by considering both the population and the context of application when testing the properties (Streiner et al. 2014). Assumptions are made from the results about the people, and their attributes, who complete the outcome measures. Therefore validity of an outcome measure is no longer seen as a simple dichotomous state of being valid or not, but more as a continual process of demonstrating validity in defined groups of people *in different situations.* For example, after testing, an outcome measure may be shown to be valid in adults *during early prosthetic rehabilitation following delivery of their primary (first) prosthesis*. This validation process looks to confirm three types of validity: content, criterion and, construct validity. These three aspects of validity are considered separately by the COSMIN group and it is advocated that they should be established independently of the other (Mokkink et al. 2010b, Mokkink et al. 2010a).

2.3.2.3.1 Content Validity

Content or face validity investigates the items or questions included in an instrument and evaluates whether it is actually looking at the construct or concept it is supposed to be when it is implemented. The content is often agreed by a group of experts when developing a new outcome measure who should clearly identify the aim of the outcome measure, the target population and the reason for selecting each of the items included. Scrutiny of the expert panel will, in part, uphold whether good content validity has been achieved. Whether individual items in a scale or sub-scales inter-correlate and are measuring the same construct or concept can be investigated through factor analyses and calculation of Cronbach's alpha, both of which give an indication of the internal consistency of a measure that is made up of more than one item / question.

2.3.2.3.2 Construct validity

Underlying factors that contribute to activities or behaviours that cannot be seen or easily measured may be referred to as hypothetical constructs. It is these constructs, or theoretical relationships that are investigated in construct validity. It is necessary to clearly list and explain the relationships that are expected, both in direction and magnitude in the form of hypotheses before performing the trial. Afterwards the results are compared to investigate if the hypotheses were proved or not.

Specificity and sensitivity are terms often used when representing the structural validity of an instrument. Specificity considers how well an instrument can detect the construct being studied, i.e. if an outcome measure has been developed to look at standing balance, measuring the person's ability or perceived ability walking up and down stairs may reflect standing balance as part of the results but the instrument may not be specific enough to measure specific aspects of standing balance ability alone. With sensitivity the ability of the measure to detect differences between different groups of patients when measuring the same construct is considered, e.g. is a measure able to detect differences between people with TTAs and TFAs, or pwLLAs in different age groups when measuring standing balance. This may also be considered as discriminative validity i.e. looking at the ability of an outcome measure to show different results for different groups. For example, the TUG times have been shown to be slower for patients aged 65 and older or for single TF amputees compared to single TT amputees, therefore the TUG has shown discriminative validity in this context (Bohannon 2006, Schoppen et al. 1999, Podsiadlo and Richardson 1991).

Convergent validity will investigate how closely related a new scale is to other measures purporting to measure the same construct. Using hypothesis testing again, the objective will be to see how the results correlate. A moderate correlation will infer that there is a relationship but not overly strong to indicate that the scales are measuring identical constructs. It is also important to look at how dissimilar results are for scales that are measuring constructs that are not related. Discriminant or divergent validity will investigate this relationship using hypothesis testing, this time looking for low correlations for the areas where they are not related.

2.3.2.3.3 Criterion validity

Criterion validity is investigated by comparing the instrument to a "gold standard" instrument. If one is not available, as is often the case in rehabilitation studies, then hypotheses can be proposed on what the results will show when compared to other similar measures. There are two main types of criterion validity; concurrent and predictive validity. Concurrent validity is investigated by giving the new measure at the same time as the "gold standard" or other comparator measure and correlations

are made. Whereas when testing predictive validity the results from the comparator measure may not be known until later in the future when the predictive quality of the new measure will be tested, again using correlation techniques.

2.3.2.4 Responsiveness

Responsiveness is considered by some to be a component of validity. However other authors believe that as there are several different elements to determining whether an instrument is sufficiently responsive it should be considered as a separate psychometric parameter. Stratford et al, (1996) asserted that responsiveness was a measure of longitudinal validity, i.e. it is a component of validity rather than a distinct entity (Stratford et al. 1996b). While Kirschner & Guyatt, 1985, stated that an effective evaluative measure has three properties; firstly reliability, which characterises low intra-subject variability, secondly validity which outlines any change detected should be consistent with an external standard and finally responsiveness which should be able to detect any clinically important changes (Kirshner and Guyatt 1985).

The terminology used is also the cause of some confusion as sensitivity to change is sometimes used instead of responsiveness. However, one distinction suggested by Liang in 1995 was that sensitivity to changes describes the ability of an instrument to measure any change, while responsiveness assesses an instrument's ability to detect clinically important change, i.e. with some context to the changes (Liang 1995). In 2000, Husted et al, attempted to bring some order to the confusion over the different terminology. In their article they highlighted the distinction between "internal" and "external" responsiveness. They also tried to clarify both the properties and interpretation of the frequently used responsiveness statistics and in addition, recommended the use of regression models in external responsiveness (Husted et al. 2000).

Internal responsiveness characterises the ability of a measure to change over a time frame. It is the ability of the instrument itself and how it is constructed to detect changes and is represented using statistical methods of analysis, e.g. effect size (ES), standardised response mean (SRM), paired t-test, Guyatt's responsiveness index (GRI), all of which present statistical significant results, and are also known as "distribution-based" methods of calculating responsiveness.

There are several different indices which measure the magnitude of change, most of which are based on ES, see Table 2.6. There is little to choose between them and the choice is likely to depend on the study design, i.e. whether there is a treatment group or not, together with a personal preference for one over the other. Cohen's effect size is considered the original and is calculated by dividing the average change by the standard deviation of the baseline scores. Guyatt's measure is used in a two group design with a pre and post treatment test. In this case it is the ratio of the average change in the treatment group to the standard deviation in the control group. The SRM is different again and is the ratio of the average change of a single group to the standard deviation of the change scores.

It should be noted that while it has been shown statistically that an outcome measure can detect a change, i.e. that the change did not occur by chance or was a measurement error, a statistically significant change may occur without the change being a clinically significant one (Husted et al. 2000). The level of "clinical significance" must take into account the patient's condition, the chronicity/stage of disease, environment etc. (Guyatt et al. 1989) and is usually defined by the clinician or expert group of clinicians, though it has been defined by the patient in some instances (Haley and Fragala-Pinkham 2006). This relates to the external responsiveness described by Husted et al (2000), as the ability of the instrument to respond to the changes in the concept it is measuring in a particular population. This can be confirmed for an outcome measure by reporting on what constitutes a change in clinical condition by either comparing results with a "gold-standard" outcome measure or by gathering data from clinicians or patients. Using another outcome measure is known as "criterion-based". Whereas gathering data from clinicians or patients is referred to as an "anchor-based" method of measuring external responsiveness and is presented using MCID, MID or MCDC, all of which present clinically important results.

Calculating the MCID is one approach to investigating how much change is enough change in an outcome measure to detect a change in the clinical condition. It was demonstrated, in a review of 29 studies that calculated 56 MCIDs, that the mean MCID was almost exactly equal to Cohen's effect size of 0.5 (moderate) (Norman et al. 2003). When considering the MCID for groups, an ES of 0.5 would be a

reasonable approximation of the threshold of important change. Three criteria were outlined for evaluating the threshold for individuals: i) the individual baseline score should be within the range found for “dysfunctional groups”, ii) the score at the end of the treatment should fall within the “normal” range, and iii) the amount of change is more than would be expected by the measurement error i.e. MDC value (Jakobsson and Westergren 2005).

Any instrument is more sensitive to large treatments than small ones and therefore it is hard to distinguish the characteristics of the measure from the characteristics of the treatment. Whichever of the methods are chosen sensitivity to change to a treatment is seen as a characteristic of both the treatment and the variance of the population being tested. Therefore the responsiveness of an outcome measure is distinct for each different clinical population unless proved otherwise.

Researchers often reported several indices of responsiveness as there is no Gold Standard or consensus on how responsiveness should be quantified or reported. Some examples are listed below in Table 2.6. The definitions in the table were taken from the following sources: (Stratford et al 1996b, Husted et al 2000, Haley and Fragala-Pinkham 2006 and Iyer et al 2003).

Table 2.6 Description of "responsiveness" indices

Responsiveness Term	Description	Measuring
Effect size (ES)	A standardised measure of change obtained by dividing the average change between baseline and follow-up measurements by the SD of the baseline,	Direction and magnitude of change
Standardised Response Mean (SRM)	The SRM is calculated by dividing the average change between initial and follow-up measurements by the SD of the change score.	Direction and magnitude of change
Paired *t* test Wilcoxon Rank Paired test	The paired t value for the difference between initial and follow-up scores is closely related to SRM. It is calculated by multiplying the SRM by the $\sqrt{(n)}$ Wilcoxon rank paired test is the non-parametric equivalent, using ranking of the pairs to calculate significance.	Significance and direction of change (but not magnitude)
Repeated Measures ANOVA Friedman's ANOVA	Similar to paired t test but used to estimate differences between several scores. Friedman's ANOVA is the non-parametric equivalent.	Significance and direction of change (but not magnitude)
Guyatt Responsiveness Index (GRI)	This index represents the ratio of observed change in a group of patients expected to undergo a change to the variability in a group of stable patients. It is calculated by dividing the change in the group expected to change by the variability in the stable patients.	Direction and magnitude of change
Minimal Clinically Important Difference (MCID)	The smallest change that would change the patient's management or be considered worthwhile or important by the clinician, caregiver, or patient	Direction, magnitude and clinical importance of change
Minimal Important Difference (MID) Minimally Clinical Detectable Change (MCDC)	Synonymous with MCID	

2.3.2.5 Interpretability of results

How an outcome measure is administrated, whether self-reported or observed will have an effect on the results. Consideration of self-efficacy and self-confidence should be taken into account as outcome measures pertaining only to the physical aspects of participation and activity also have to consider psychosocial factors that may be an influence especially in those outcome measures that rely on the individual to score themselves in self-reported outcome measures. Therefore, consideration of the population under study, knowledge of the choice of statistical analysis method used to investigate it's psychometric properties and how it was administered are critical to understanding whether the outcome measure is an appropriate choice or not.

Most research and clinical studies collect data in groups and understanding whether the results obtained can be interpreted at the individual level is also key. When results are presented following repeated measurements of groups of participants / patients they are presented as reaching an acceptable standard and consequently the results can then be interpreted for other similar groups, or at the group level. However, closer inspection of how the results were calculated or more stringent standards may be required if the results are to be interpreted for individual patients, or at the individual level. The identification of whether results can be interpreted at an individual level is not always presented explicitly and may be inferred or assumed by the reader or clinician. For example ICC values should be at least 0.90, and preferably higher than 0.95, for the results to be acceptable at the individual level, whereas an ICC or 0.7 is acceptable if the outcome measure is used when measuring a group of patients (Shrout and Fleiss 1979, Landis and Koch 1977, Cicchetti 1994). Further discussion on this aspect will be presented throughout the following chapters.

2.3.3 International Classification of Functioning Disability and Health

The ICF framework was developed by the World Health Organisation (WHO). It is a framework that sets out to define components of health and well-being (USTUN et al. 2003). The framework uses a common language and the intention is to facilitate communication across the world through the various health professional disciplines about health status and health care. Use of these common terms when articulating

the constructs being measured by different outcome measures will assist when comparing or combining results of outcome measures that measure the same constructs.

The components of the framework address the main areas of Functioning and Disability with other contextual factors also being taken into account. There are two component areas which are further sub-divided: a) Body and, Functions and Structure, where body functions are defined as the physiological (including the psychological) functions, and body structures are defined as the anatomical parts; b) Activities and Participation, where any task or action performed by an individual as part of their everyday life is considered an activity and limitations are considered any difficulties the individual may have undertaking these activities. The contextual factors which are considered are: environmental i.e. factors external to the individual but may have an influence and; personal factors e.g. gender, age, lifestyle habits etc., all of which are not classified in ICF but are collected. The relationships can be seen in Figure 2.1

The difference between performance and capacity is subtle but important within this classification. Within the ICF framework, if an individual has a problem executing a task or action they are classified as having a limitation at the person level.

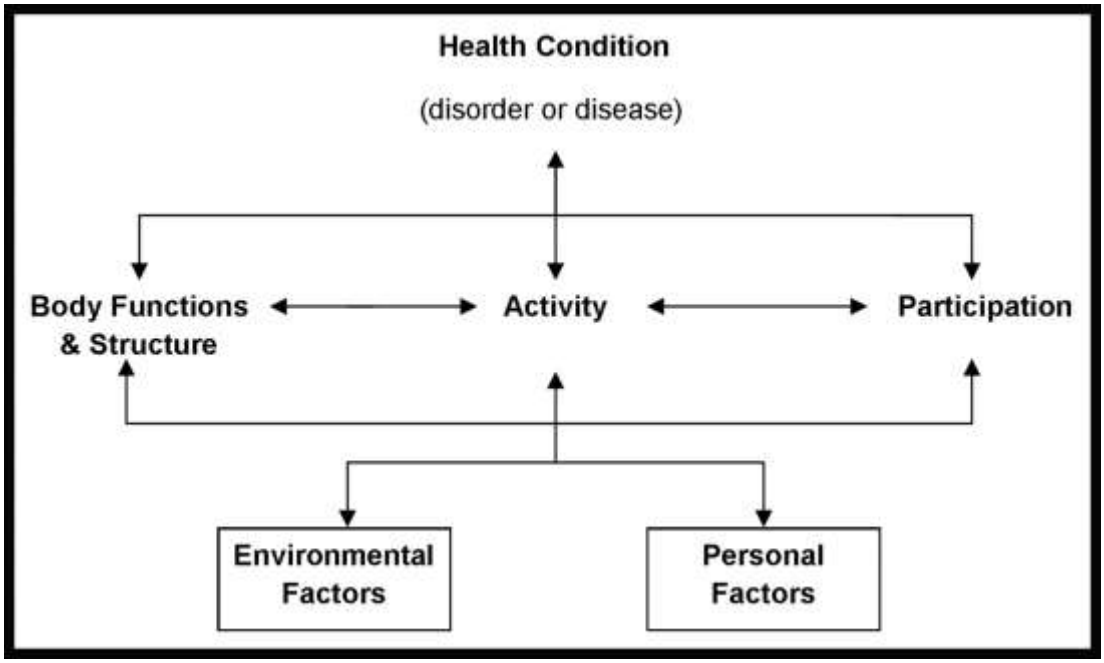


Figure 2-1 ICF framework (WHO, 2002)

It is important to consider what an individual is able to do within their current environment, including the use of assistive devices or support if required as this is measured as their “performance”. However, if, for example, an amputee is assessed without their prosthesis and walking aids (if required), their capability or “capacity” in this “standardised environment”, is likely to be reduced. If information is known about both the “capacity” and “performance” it will be clear if the person's current environment has enabled him/her to perform better than their “capacity” would predict. This will also help understand whether some aspect of the environment is a barrier to their performance. This difference between “performance” and “capacity” is an important aspect of the ICF classification but it is one that is rarely captured by the use of outcome measures that only consider performance of pwLLAs. Self-reported outcome measures may capture some of this subtlety but it is not known what influence self-efficacy and self-confidence has on the results. This has not been considered and is outside the scope of this thesis, however is worthy of consideration in future work.

Rehabilitation impacts on activity for an individual however, as seen from the bio-psychosocial model of disability that is the basis for ICF, activity at an individual level is interlinked with impairments at the body level and also limitations at a societal level. An individual is considered to have limitations at a societal level if they are unable to perform a task to participate or be involved in personal, social or work activities. For a pwLLA the use of either a wheelchair or a prosthesis, as well as any underlying issues with their body function, will influence their activity patterns. These will then have an impact on the level they interact in society on personal, social and economic terms.

An instrument called the “Rehabilitation Problem-Solving Form” (RPS-Form) has been developed which allows health care professionals to assess patient problems, particularly from the patients’ perspective, from within the ICF framework (Steiner et al. 2002). Outcome measures that are based on the ICF have the potential to improve the recording, monitoring and communication of a patients’ rehabilitation and also their limitations. Numerous ICF Core Sets have been published for a number of conditions (Steiner et al. 2002, Cieza et al. 2004). A recent study by Kohler et al, (2011) has looked at the feasibility of using a condition-specific Core Set for pwLLAs as the basis for an outcome measures. Initial findings show the

Core Set has content validity and sensitivity as a measure of change in patients following an amputation (Kohler et al. 2009, Kohler et al. 2011). However, further work is required to demonstrate its psychometric properties, as well as promote its use in clinical practice.

In an effort to evaluate which outcome measures, currently used with pwLLAs, have been classified within the ICF, two systematic reviews were carried out. One explored the ICF category of activity and participation and identified a number of outcome measures that had the potential for use during the rehabilitation period following amputation (Deathe et al. 2009). The other review looked at outcome measures in the category of body function or structure, but identified only a few measures that had been well validated (Hebert et al. 2009). The authors of both reviews agreed that responsiveness of any of the outcome measures used to detect changes in the amputee's performance or condition had not been well established. They recommended that this be the focus of future studies, along with the continued establishment of validity and reliability. Authors of a previous review had concluded that while some outcome measures were valid, the studies included in the review were complex using "inconsistent" and "confusing terminology" which clinicians would not find helpful when choosing an outcome measure (Condie et al. 2006).

Two more reviews looking at outcome measures used with pwLLAs have been published recently (Hawkins et al. 2014, Heinemann et al. 2014). Hawkins et al (2014) did conclude that there were a number of "high quality" validated instruments. However, the quality rating criteria used (based on Johnston and Graves 2008) described the level of evidence for these "high" quality outcome measures as "adequately/reasonably valid for the main defined purpose" and "OK to use in studies, although checking of assumptions or small improvements may be desirable to further improve the measure (e.g. classical measures would benefit from item-response theory or Rasch analysis)". The narrative review presented by Heinneman et al (2014) did not contain any quality assessment of the studies reviewed, but the authors did note that they were encouraged that responsiveness of measures is often reported. However, only SEM or MDC values were presented as evidence of responsiveness or sensitivity within the summary table, with no other indices of responsiveness provided to support the claims. The last of their key points was that "emerging information about responsiveness of outcome measures improves their

clinical utility", but there is clearly still more work required to provide this in a form that is understandable for clinicians in the field of prosthetic rehabilitation.

There is a wide variety of outcome measures purporting to measure aspects of physical function, and a full understanding of their psychometric properties will help clinicians and researchers choose the most appropriate. However, the apparent lack of evidence available on these psychometric properties, has led to the second, third and fourth research questions in this thesis:

Research Q2
Which outcome measures, used to measure physical function during prosthetic rehabilitation, have been investigated and published in peer-reviewed journals; and what are the characteristics (i.e. psychometric properties) that have been presented?

Research Q3
What is the methodological quality of the published studies presenting results for psychometric measurement properties of the outcome measures used to measure physical function during prosthetic rehabilitation?

Research Q4
What are the levels of evidence presented in these studies?

These questions will be addressed in Chapter 4.

Providing more data on the psychometric properties of outcome measures commonly used by AHPs in prosthetic rehabilitation is likely to be of value to both clinicians and researchers in this field. Having an understanding of the reliability, measurement error and responsiveness is especially important when using outcome measures to evaluate changes following an intervention. However, a lack of such information on outcome measures used with pwLLAs has been commented upon in several systematic reviews (Kannenberg et al. 2014, Fortington et al. 2012, Deathe et al. 2009, Hebert et al. 2009, Orendurff 2012, Shultz and Olszewski 2013).

This need for more data on reliability, measurement error and responsiveness of outcome measures has prompted the final two questions in this thesis:

Research Q5
What are the reliability and measurement error parameters of outcome measures most regularly used by AHPs for the assessment of physical function during prosthetic rehabilitation?

This question will be addressed in Chapter 5

Research Q6
What is the responsiveness of Physical Function outcome measures regularly used with lower limb amputees when assessing change during the rehabilitation period immediately following limb fitting?

This question will be addressed in Chapter 6

2.4 Measuring the quality of studies presenting psychometric data

Standardised checklists are available that check the methodological quality of studies that investigate the impact of treatments and interventions on patients. However, the quality of any outcome measure used within these studies is considered only with a yes/no check, on whether the measure(s) used were “reliable” and/or “valid” and/or “responsive”. Studies presenting data on the psychometric measurement properties of outcome measures that measure physical function of pwLLA’s have been investigated in several reviews (Rommers et al. 2001, Condie et al. 2006, Deathe et al. 2009, Hebert et al. 2009, Hawkins et al. 2014). The most commonly used criteria used to measure the quality of the studies in these reviews were published by Johnston & Graves in 2008 (Johnston and Graves 2008). These criteria outline aspects of reliability and validity that should be presented when developing or testing a measure in a new population. An overall

quality rating is given taking into account the absence or presence of evidence of reliability and/or validity. However, the level of the evidence presented per measurement property for each outcome measure is not considered.

In 2007, Terwee et al published criteria that could be used when measuring the quality of studies by individually rating the validity, reliability and responsiveness, as well as the interpretability of results (Terwee et al. 2007). This was part of the early work by the COSMIN group as they worked to bring together a consensus of opinion on how measurement property parameters should be defined, studied and reported. The primary aim stated on their website is: “improve the selection of health measurement instruments”. As part of this initiative, the COSMIN group developed a critical appraisal tool or checklist that contains standards for evaluating the methodological quality of studies on the properties of health measurement instruments (Mokkink et al. 2010a, Mokkink et al. 2010c). It was developed as part of an international Delphi study as a multidisciplinary, international collaboration (Mokkink et al. 2006). While the focus is on HR-PROs, the checklist can also be used for evaluating studies on other kinds of health measurement instruments, such as performance-based instruments or clinical rating scales. It has been updated since the first release and now rates the items within each measurement property on a 4-level scale. It has been used in many systematic reviews since it has been published (e.g. Dobson et al. 2012, Pin 2014, Green et al. 2014, Ammann-Reiffer et al. 2014, Larsen et al. 2014, Proud et al. 2015).

2.5 Rationale for the overall aim(s) of the thesis

The issues of the variability and errors associated with measurement have been introduced in this background chapter to highlight the importance of establishing the “robustness” and “appropriateness” of any outcome measure in use. This is specifically important for pwLLA where provision of correct prosthetic component has a positive impact on their limb-wearing function. There is little consensus on prosthetic provision guidelines, beyond the initial price of the component and classification of activity in the early stages of rehabilitation. This is potentially limiting for amputees, particularly for those with lower activity ability as they often receive little beyond the basic componentry. The need to test more technologically-advanced components for effectiveness in this population is critical. However, there

is still little agreement on which outcome measurement tools to use. More importantly, the evidence of their psychometric qualities is lacking so the choice for clinicians about which are appropriate and robust enough to effectively measure the effect of interventions is not clear.

This background chapter has outlined the need for improved understanding of which outcome measures are regularly being used with pwLLAs, (research question1) because of the continued lack of a consensus of practice, with regard to the collection of outcome measurement data in this population. A survey conducted with clinical staff will attempt to provide some information in this area (chapter 3).

It has also been demonstrated that there are still gaps in the evidence to support the use of specific physical outcome measures during prosthetic rehabilitation. Therefore the need to have a clearer understanding of what the current quality of evidence of the psychometric properties of the outcome measures used with lower limb amputees, (research questions 2, 3 and 4) will be addressed by a systematic review of the literature (chapter 4).

The evidence for reliability, measurement error and responsiveness of these outcome measures is particularly lacking. Therefore, there is a need to add to this evidence by providing a clearer understanding of the reliability, measurement error and responsiveness of the outcome measures currently used in clinical practice (research questions 5 and 6). Both the test re-test study (study III in chapter 5) and the longitudinal cohort study (study IV in chapter 6) will address these gaps.

This thesis will therefore attempt to answer each of the research questions to help provide a better overall understanding of the psychometric properties of outcome measures used to measure changes in physical function in pwLLAs during prosthetic rehabilitation. An overview of the thesis, visually presenting the relationship between the research questions and the gap in the current evidence is presented in Figure 2.2.

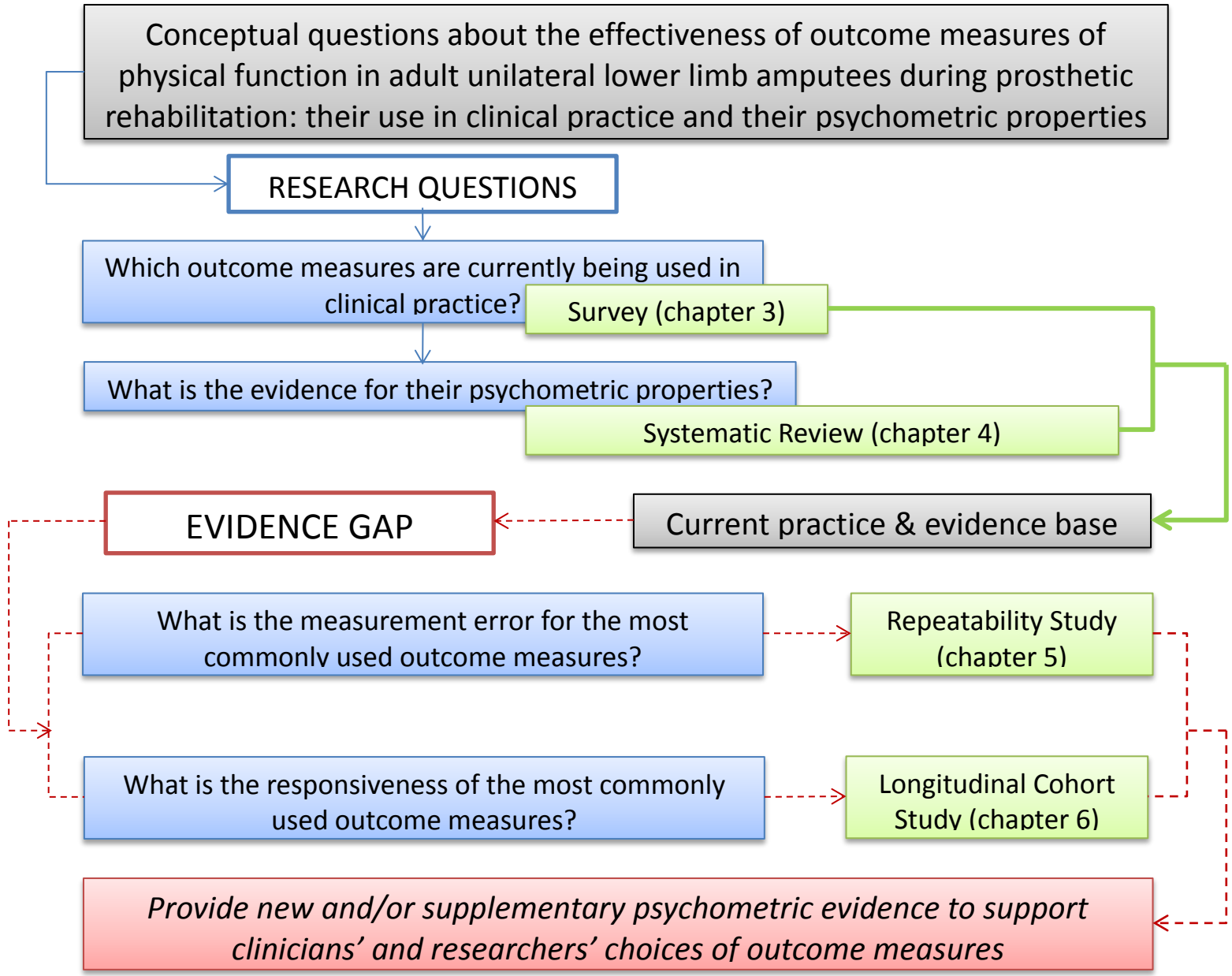


Figure 2-2 Thesis overview

3 Chapter 3 Survey (study I)

3.1 Purpose of chapter

The purpose of this chapter is to describe the methodology, and present and discuss the results of a survey carried out with Allied Health Professionals (AHP's) who work in prosthetic rehabilitation across the UK. The survey was carried out to answer Research Question 1.

Research Q1 What outcome measures are used regularly in clinical practice during prosthetic rehabilitation?

A supplemental question was also considered when designing the survey: Which factors are considered influential in contributing to a successful prosthetic outcome for a pwLLA from both the AHP and pwLLA's perspective?

3.2 Background

3.2.1 Outcome measurement with lower limb amputees

The large choice of outcome measures that can be used with pwLLAs has been highlighted in several reviews of the literature (Rommers et al. 2001, Condie et al. 2006, Deathe et al. 2009, Hebert et al. 2009, Hawkins et al. 2014, Heinemann et al. 2014, Sinha and Wim J. 2011). However, the number of outcome measures identified in these reviews, suggests that there is a lack of consensus about which outcome measure(s) should be used, under which circumstances, and with which patients.

Turner-Stokes & Turner-Stokes surveyed members of the British Society of Rehabilitation Medicine (BSRM) Consultants in 1997, regarding their use of outcome measures in all areas of rehabilitation medicine. Only the Harold Wood score (now known as the SIGAM) was listed, by respondents to the survey, as an outcome measure designed specifically for use with pwLLAs (Turner-Stokes and Turner-Stokes 1997). When Skinner and Turner-Stokes repeated the survey in

2005, no other amputee-specific outcome measures were included, but a growing usage of the SIGAM was noted (Skinner and Turner-Stokes 2006). This increased use of the SIGAM may have been in response to the publication of a study by Ryall et al in 2003, reporting on its reliability, validity and responsiveness. The results from the 2005 survey by Skinner and Turner-Stokes (2006) were used to update the BSRM "Basket" of Measures recommended for their members, and consequently the SIGAM was included.

Increased pressure to use valid and reliable outcome measures in clinical practice from Professional Bodies in the late 1990's and early 2000's (Skinner and Turner-Stokes 2006, Hammond 2000, Moseley et al. 2002), appears to have prompted the release of some documents. In addition to the BSRM recommendations, the British Association of Chartered Physiotherapists in Amputee Rehabilitation (BACPAR) produced an Outcome Measures Toolbox to assist clinicians with the choice of which outcome measure to use during prosthetic rehabilitation. To be included in the Toolbox, outcome measures were required to have evidence of their validity, reliability and responsiveness, as well as being practical to use. The BACPAR Toolbox is intended to be a guide for clinicians and is not mandatory; the decision regarding when to use the outcome measure, i.e. which patient group, and at which time point during the rehabilitation period, is left to the clinician. The Toolbox was updated in 2014 and the currently recommended outcome measures are: the Activities-specific Balance Confidence (ABC) Scale-UK, Amputee Mobility Predictor (AMP), Houghton Scale, Locomotor Capabilities Index (LCI)-5, Trinity Amputation and Prosthesis Experience Scales (TAPES), Timed Up and Go (TUG), L-Test, 2min and 6 min timed walk tests and the Berg Balance Scale (BBS) (Cole et al. 2014).

More recently the British Association of Prosthetists and Orthotists (BAPO) released a document, containing advice for their members on the use of outcome measures (Young et al. 2015). It included the recommendation to use outcome measures, where appropriate, routinely in their clinical practice and includes: the Numeric Rating Scale (NSR-11) for pain; the TUG and 10 metre walk test, to be used for lower limb patients, as well as the Socket Comfort Score (SCS) exclusively for use with amputees.

Although lists of recommended outcome measures are available, it is not known which, if any, are regularly used in clinical practice. Nor is it known whether there is a difference in the outcome measures being used with pwLLAs of different activity levels. Therefore, the primary aim of the survey was to establish which outcome measures are being regularly used by AHP's, i.e. physiotherapists, occupational therapists and prosthetists during prosthetic rehabilitation.

3.2.2 Factors influencing prosthetic outcome

There are several physical factors, such as balance, pain, and muscle strength, which play a role during prosthetic rehabilitation and have an influence on the final outcome. The final outcome is unique to each amputee, but a prosthetic outcome may be considered successful when the amputee is mobilising maximally with their prosthesis. Psychosocial factors also have an impact on the prosthetic outcome. Sansam et al (2009) reviewed the literature to investigate which factors predicted walking with a prosthesis (Sansam et al. 2009). They concluded that comparisons were difficult across the identified studies and there were some conflicting conclusions with regard to positive predictive factors (Sansam et al. 2009). In order to identify which factors were considered by pwLLAs to be most impaired, Demet et al (2003) carried out a study with Part I of the Nottingham Health Profile (NHP). The NHP is used to measure perceived distress associated with a variety of health conditions and Part I assesses the level of distress in 6 areas: energy levels; pain; emotional reactions; sleep; social isolation; and physical disability. Respondents perceived that energy levels, pain and physical disability were the areas that were impaired the most (Demet et al. 2003). However, it was not clear if these were perceived to be most important to the pwLLAs, in terms of their rehabilitation.

A secondary aim of these surveys therefore, was to find out which factors both AHP's and pwLLAs, considered most influential in contributing to a successful prosthetic outcome.

3.3 Methodology

3.3.1 Survey questionnaire design

Two survey questionnaires were designed specifically for use in this study, one for AHP's and one for pwLLAs, (see appendices 1 & 2). Although the surveys were answered anonymously, both included questions to obtain basic demographic variables. In the AHP survey, the respondents were asked to state: their profession; the length of time they had been qualified; the length of time they had worked with amputees; and the amount of time per week they worked in prosthetic rehabilitation. The survey for pwLLAs asked for their age-range, sex, level of amputation and number of years of prosthetic use. Both surveys were piloted within a multi-disciplinary prosthetic team. The team also asked patients currently undergoing rehabilitation to feedback on the patient survey. Minimal comments on the readability and appropriateness of the questions were received and only minor changes were made.

3.3.1.1 Inclusion criteria for pwLLA respondents

The pwLLA had to meet the following inclusion criteria: established limb wearers of at least one year duration; and attending a prosthetic clinic. Established limb wearers were those who had been wearing their prosthesis for at least one year, were wearing it for at least eight hours per day and could walk at least one kilometre outdoors, with or without a walking aid. These criteria approximately equate to level K3 and above on the Medicare K-Level classification system (see Table 3.1).

Table 3.1 K-Level classification (K3 & K4 - Adapted from AAO&P website)

K-Level	Description
K3	The patient has the ability or potential for ambulation with variable cadence with the ability to traverse most environmental barriers and may have vocational, therapeutic, or exercise activity that demands prosthetic use beyond simple locomotion
K4	The patient has the ability or potential for prosthetic ambulation that exceeds basic ambulation skills, exhibiting high impact, stress, or energy levels

Suitable patients were introduced to the survey by their prosthetist and all surveys were completed and returned anonymously to the researcher in stamped addressed envelopes.

3.3.1.2 List of Outcome Measures in AHP survey

The AHP survey included a list of 30 outcome measures which are validated for use with pwLLAs (see table 3.2). The list included both patient-reported outcome measures (PROMs) where a respondent either completes a questionnaire themselves or has his/her answers recorded, and observed measures, where an observer (clinician or researcher) records the activity of an amputee, either using a criterion-based rating scale or direct performance scores, i.e. distance or time. There were also outcome measures that used a combination of data collection methods included.

Table 3.2 Outcome measures included in AHP Survey

Outcome Measure	Type	Main Constructs being Measured
Activities-specific Balance Confidence (ABC) Scale-UK	PROM	Balance, Functional Mobility
Amputation Related Body Image Scale	PROM	Body Image
Amputee Activity Score	PROM	Functional Activities
Amputee Mobility Predictor with Prosthesis	Observed	Functional Activities
Attitude to Artificial Limb Questionnaire	PROM	Satisfaction with prosthesis
Barthel Index	Mixed	Activities of Daily Living (ADL)
Body Image Questionnaire	PROM	Body Image
Canadian Occupational Performance Measure (COPM)	Mixed	Functional Activities
Frenchay Activities Index (FAI)	PROM	Functional Mobility, ADL, Occupational performance
Functional Measure for Amputees (FMA)	PROM	Functional Activities
Functional Independence Measure (FIM)	Observed	Functional Activities
Houghton Scale	PROM	Prosthetic Use
Locomotor Capabilities Index (LCI) or Modified LCI-5	PROM	Functional Mobility
Orthotics and Prosthetics National	Mixed	Health-related Quality of Life

Outcomes Tool (OPOT)		(HR-QoL), Functional Mobility
Patient Generated Index	PROM	ADL, General Health, Motor Activities
Perceived Social Stigma Scale	PROM	NR-QoL
Prosthesis Evaluation Questionnaire (PEQ)	PROM	Prosthetic-related Quality of Life (PR-QoL)
Prosthetic Profile of the Amputee (PPA)	PROM	PR-QoL, Functional Activities
Prosthetic Observational Gait Score (POGS)	Observed	Gait
Rivermead Mobility Index (RMI)	Mixed	Functional Mobility, Balance
Russek's Code	Observed	Functional Ability
Short Form 12 or 36 Health Survey (SF-12 or 36)	PROM	HR-QoL
Sickness Impact Profile (SIP)	PROM	HR-QoL, ADL
Special Interest Group in Amputee Medicine (SIGAM) Mobility Grades	PROM	Functional Mobility
Socket Comfort Score (SCS)	PROM	Pain / Discomfort
Timed “Up and Go” Test (TUG)	Observed	Mobility and Balance
Timed Walk Test (TWT) - 2min (2MWT) or 6min (6MWT)	Observed	Mobility Endurance
Trinity Amputation and Prosthesis Experience Scales (TAPES)	PROM	HR-QoL

The AHP respondents were asked to indicate which, if any, of the outcome measures listed, they used regularly with low, medium and high level activity amputees. The amputee’s activity level was defined by their K-Level, see Table 2.1 in chapter 2 for full list of K-level criteria. In addition, AHP’s were asked to indicate whether they use the outcome measures with primary and/or established limb-users. Primary limb-users are considered to be those who have just received their primary or first limb, while established limb-users are those who are considered, by their AHP, to be established on their primary limb, or be at least one year post primary limb delivery. Information was also gathered about whether the respondents had stopped using any of the outcome measures, and if so why. Finally, they were also asked to name any outcome measures they used regularly but which were not in the included in the original list of 30.

3.3.1.3 Factors influencing the outcome of prosthetic rehabilitation

Both surveys included a list of ten factors that could potentially influence a successful outcome following prosthetic rehabilitation. The AHP respondents were asked to consider the factors with respect to outcomes according to each K-level (see Table 2.1). Amputee respondents were asked to consider the factors with respect to their own prosthetic outcome. According to the selection criteria, this was equivalent to at least a K3 level, i.e. that of a typical community limb-user whose activities extend beyond simple ambulation (see Table 3.1). The factor list was compiled and adapted using clinical experience. The findings from the Demet study (Demet et al. 2003), where respondents perceived energy levels, pain and physical disability were used as a starting point for a draft list. Some level descriptors and dimensions from the International Classification of Functioning, Disability, and Health (World Health Organisation September 2003) were also used to add to the draft list. For example: from Body Functions – "b110 Consciousness functions (state of awareness and alertness)" is associated with motivation; from Activities and Participations – "d410 & d415 Changing & maintaining body position" relate to balance and coordination; and in Environment - "e3 Support and relationships" includes family and friends. The draft list became the final list, see (Table 3.3) after it was piloted with a group of pwLLAs and AHPs and no additional factors were suggested.

Table 3.3 Factors influencing a positive prosthetic outcome.

Factor	**Example ICF codes and descriptors**	
Balance & Coordination	d410 & d415	Changing & maintaining body positions
Energy, Drive & Motivation	b110	Consciousness functions (state of awareness and alertness)
General Muscle strength	b730	Muscles power functions
General Range of Movement	b710	Mobility of joint functions
Manual Dexterity	b710 & b730	Specifically of the hand and fingers
Memory	b144	Memory functions
Pain Control	b280	Sensation of pain
Self Confidence	b1266	Confidence
Socket fit & comfort	b280	Sensation of pain
Support of family & friends	e3	Support and relationships

All respondents were asked to rank, in order of importance from 1 to 10, with '1' being the most important, the factors they considered the most important to achieving successful prosthetic rehabilitation. They also had the opportunity to list any additional factors they felt were important.

3.3.2 Survey distribution

The AHP survey was distributed during May, June and July 2013.

Surveys were sent via professional networks and sub-groups where possible. For example: surveys were sent to physiotherapists, by e-mail, to all BACPAR members (n=137) as well as to all members of the Scottish Physiotherapists in Amputee Rehabilitation Group (SPARG) (n=27). This was done through the network secretaries who sent an e-mail to all members.

The survey was also distributed to occupational therapy colleagues through their Special Interest Groups both in England and Scotland, but it was through one of the members and not the network secretary and therefore the sample size was unknown.

Distribution amongst prosthetists was initially less widespread, as the survey was only e-mailed to the Lead Prosthetists at the three largest centres in Scotland (Dundee, Edinburgh and Glasgow) asking them to distribute throughout their teams. Onward distribution by physiotherapists to colleagues within prosthetic multi-disciplinary teams to both prosthetists and occupational therapists helped to encourage as wide a reach as possible, within the limited contacts of the researcher.

The multi-disciplinary nature of prosthetic rehabilitation with a relatively small number of specialist centres (n=43 across the UK) meant that some clinicians may have received the survey several times. All surveys were completed anonymously and returned to the researcher using either stamped addressed envelopes or by e-mail. If any surveys were returned by e-mail, they were saved and the e-mails deleted prior to the data being collated, to protect the anonymity of the respondents.

3.3.3 Statistical analysis

Descriptive analysis was undertaken detailing the characteristics of the respondents, and summarising the ranking of the factors in order of importance and the use of outcome measures commonly used by AHP's.

The rankings of factors listed were averaged for each of the two groups of respondents, i.e. the AHP's and the pwLLAs, to establish an overall ranking for each group. In addition, the responses from AHP's for K3 and K4 were averaged to match the activity levels of the pwLLA respondents.

3.3.4 Ethical opinion

Prior to conducting the surveys, the Scientific Officer for the South East Scotland Research Ethics Service was contacted. They advised that an ethical review was not required under the terms of the Governance Arrangements for Research Ethics Committee in the UK. The project was considered an opinion survey seeking the anonymous views of NHS staff and patients on aspects of service delivery.

3.4 Results

3.4.1 Allied Health Professional respondent characteristics (or descriptors)

By the end of the three month recruitment period, 23 physiotherapists, seventeen prosthetists and five occupational therapists had responded (see Table 3.4). These 45 AHP's represented a convenience sample following a pragmatic decision to close the survey at the end of three months. No more responses were received after this period. The majority (34/45) of respondents had been qualified for over 10 years and the same percentage (76%) had been working with amputees for at least 5 years. Almost three quarters (33/45) of the respondents worked with amputees for at least two-thirds of their working week.

Table 3.4 Characteristics of AHP respondents

Professional Group	Years qualified			Years working with amputees			% of work connected with amputee rehabilitation		
	< 5	5-10	>10	< 5	5-10	>10	0 - 30	31 - 60	61 – 100
Physiotherapists n= 23	1	2	20	5	5	13	4	5	14
Prosthetists n=17	3	2	12	2	3	12	2	0	15
Occupational Therapists n=5	1	2	2	2	2	1	1	0	4
Total n=45	5	6	34	9	10	26	7	5	33

3.4.1.1 Response Rates

Assuming all physiotherapists were members of BACPAR (n=137, at time of survey) this represents a 17% (23/137) return rate. To protect the anonymity of respondents, information on their place of work was not collected. However, 14 of the physiotherapists declared that at least two-thirds of their time was connected to amputee rehabilitation. Prosthetic centres across the UK, each employ a multi-disciplinary team (MDT) to assess and deliver the specialist prosthetic rehabilitation. Assuming that each of the physiotherapist respondents represented a prosthetic centre, of which there are 43, the return rate rises to 32% (14/43). The rate was 44% (19/43) with the inclusion of physiotherapists who worked for more than a third of their time in amputee rehabilitation.

It is difficult to estimate a return rate for prosthetists on the same basis. They are dual trained in both Prosthetics and Orthotics, and their professional body, BAPO, includes a large number of practising orthotists who do not see pwLLAs.

3.4.2 Lower limb amputee respondent characteristics (or descriptors)

Twelve pwLLAs had responded by the end of July 2013 (see Table 3.5 for details). Four were aged between 26 and 50 years with the remaining eight over 51. Eight were unilateral amputees, with only one at trans-femoral level. Two were bilateral at

the trans-tibial level and two had other level amputations. The majority (eight) had their amputation over 10 years ago.

Table 3.5 Characteristics of pwLLA respondents

Age range				Gender		Level (Bilateral)				Years had a prosthesis		
18-25	**26-50**	**51-65**	**>65**	**M**	**F**	**TTA**	**TKA**	**TFA**	**Other**	**1-5**	**5-10**	**>10**
0	4	4	4	8	4	7 (2)	0 (0)	1 (0)	2	3	1	8

3.4.3 Outcome measures commonly used in clinical practice

Thirty two of the 45 AHP respondents (71%) indicated they used outcome measures regularly, i.e. at least once a week. Breakdown by professional group was: 100% of physiotherapists (23/23), 41% of prosthetists (7/17) and 60% of occupational therapists (3/5).

Physiotherapists indicated that they used a total of 14 different outcome measures from the list. The prosthetists used 8 different outcome measures, though only the SIGAM and SCS most regularly. The occupational therapists used only three. Table 3.6 illustrates the numbers of all AHP respondents who regularly used the outcome measures listed in the survey and at what level. The choices of the physiotherapists, prosthetists and occupational therapists respondents are detailed in tables included in the Appendix 3.

Table 3.6 Total number of AHP respondents who regularly used the listed outcome measures

	Activity Level K0 – K2 Non-limb wearer, indoor or limited outdoor ambulator		**Activity Level K3** Wearing limb daily and fully ambulant outdoors		**Activity Level K4** Undertakes athletic activities in addition to daily tasks		**TOTALS**
	Primary	**Established**	**Primary**	**Established**	**Primary**	**Established**	
Activities-specific Balance Confidence (ABC) Scale-UK	3	1	3	2	1	2	**12**
Amputee Activity Score	3	3	3	3	2	3	**17**
Amputee Mobility Predictor with Prosthesis	4	2	4	2	2	1	**15**
Attitude to Artificial Limb Questionnaire	1	0	1	0	0	0	**2**
Barthel Index	3	0	2	0	2	0	**7**
Canadian Occupational Performance Measure (COPM)	2	1	2	0	0	0	**5**
Functional Independence Measure (FIM)	2	0	2	0	1	0	**5**
Houghton Scale	5	4	5	4	1	1	**20**
Locomotor Capabilities Index or LCI-5	11	8	13	9	7	6	**54**
Prosthesis Evaluation Questionnaire	1	1	1	2	0	2	**7**
Short Form 12 or 36 Health Survey (SF-12 or 36)	1	0	0	4	0	0	**5**
Special Interest Group in Amputee Medicine (SIGAM) Mobility Grades	16	14	16	13	9	10	**78**
Socket Comfort Score (SCS)	9	10	9	10	6	7	**51**
Timed “Up and Go” Test (TUG)	14	12	13	11	7	6	**63**
Timed Walk Test - 2min or 6min (TWT)	10	11	12	12	6	6	**57**
Trinity Amputation and Prosthesis Experience Scales	2	2	3	6	2	2	**17**

The totals in the last column exceeds the total number of respondents because they could select each outcome measure for more than one activity level with both primary and established amputees.

The five outcome measures used more often than any others were: SIGAM Grades (78 out of maximum possible 270 [6 options x 45 respondents]), TUG (63), TWT (57), LCI / LCI-5 (54) and SCS (51). Two of these are observed performance measures; the TUG which measures mobility and balance (Podsiadlo and Richardson 1991) and the TWTs which measures mobility and endurance (Butland et al. 1982). The other three are PROMs; the SIGAM (Ryall et al. 2003b) and LCI / LCI-5 (Gauthier-Gagnon et al. 1998, Franchignoni et al. 2007b) both measure functional mobility, and the SCS (Hanspal et al. 2003) measures the patient's perceived comfort of his/her socket. Respondents also identified that the use of several outcome measures had been discontinued (see Table 3.7). Various reasons were given and in no particular order: lack of time to collect data; no relevance to clinical practice; lack of sensitivity i.e. ceiling and/or floor-effect observed, took too long to administer, not user or therapist-friendly or a license was required.

Table 3.7 Outcome measures discontinued and/or not selected by AHP respondents

Outcome Measure	**Type**	**Main Construct being Measured**	**Discontinued (D) and / or Not Selected (NS)**
Activities-specific Balance Confidence Scale-UK	PROM	Balance, Functional Mobility	NS
Barthel Index	Mixed	Activities of Daily Living (ADL)	D
Body Image Questionnaire	PROM	Body Image	NS
Frenchay Activities Index (FAI)	PROM	Functional Mobility, ADL, Occupational performance	NS
Functional Independence Measure (FIM)	Observed	Functional Activities	D
Functional Measure for Amputees (FMA)	PROM	Functional Activities	D and NS
Houghton Scale	PROM	Prosthetic Use	D
Locomotor Capabilities Index (LCI) or Modified LCI-5	PROM	Functional Mobility	D
Orthotics and Prosthetics	Mixed	Health-related Quality of	NS

National Outcomes Tool (OPOT)		Life (HR-QoL), Functional Mobility	
Patient Generated Index	PROM	ADL, General Health, Motor Activities	NS
Perceived Social Stigma Scale	PROM	NR-QoL	NS
Prosthesis Evaluation Questionnaire	PROM	Prosthetic-related Quality of Life (PR-QoL)	D
Prosthetic Profile of the Amputee (PPA)	PROM	PR-QoL, Functional Activities	D and NS
Prosthetic Observational Gait Score (POGS)	Observed	Gait	NS
Rivermead Mobility Index (RMI)	Mixed	Functional Mobility, Balance	NS
Russek's Code	Observed	Functional Ability	D and NS
Sickness Impact Profile (SIP)	PROM	HR-QoL, ADL	NS
Special Interest Group in Amputee Medicine (SIGAM) Mobility Grades	PROM	Functional Mobility	D
Trinity Amputation and Prosthesis Experience Scales (TAPES)	PROM	HR-QoL	D

While thirteen of the outcomes measures were not selected, fourteen additional outcome measures were named when the respondents were asked to list any other they used regularly, see Table 3.8. One respondent also reported using their own questionnaire, though no further details were given.

Table 3.8 Additional outcome measures listed by AHPs

Outcome Measure	Type	Main Construct being Measured
Clinical Test of Sensory Integration and Balance (CTSIB)	Observed	Balance
EQ-5D-5L™	PROM	Global Health
Functional Goal Setting	PROM	Functional Activities
Goal Attainment Scale (GAS)	PROM	ADL, General Health, Motor Activities
Hospital Anxiety and Depression Score (HAD)	PROM	Anxiety and Depression
L-test	Observed	Mobility and Balance
Reintegration into Normal Living Index (RNLI)	PROM	ADL, Social Relationships
Satisfaction with Prosthesis Score	PROM	Satisfaction with Prosthesis

(SATPRO)		
Step Test	Observed	Mobility and Balance
4-step Square Test	Observed	Mobility and Balance
T-test	Observed	Mobility and Balance
Trans-femoral Predictor Tool	Observed	Functional Activities
Visual Analogue Scale	PROM	Any – depends on the question asked
Walking Ability Questionnaire (WAQ)	PROM	Functional Mobility

3.4.4 Important factors

The ranking of factors influencing a successful rehabilitation chosen by AHP's is split by activity level and whether the patient is a primary or established amputee. However, not all AHP respondents indicated a ranking for each activity level possibly indicating a lack of experience with amputees at this activity level, or they don't use OM for that activity level. Slightly fewer responses in the "established" columns may reflect a similar lack of experience, but also a lack of contact with AHP's beyond the "primary" rehabilitation stage. Details of all responses are shown in Table 3.9.

There was only one additional factor added by any of the respondents and that was considered to have any influence: "the relationship between the amputee and the AHP".

Table 3.9 Average ranking of factors influencing a successful rehabilitation by AHP's

Factor	**Activity Level K0 – K2** Non-limb wearer, indoor or limited outdoor ambulator		**Activity Level K3** Wearing limb daily and fully ambulant outdoors		**Activity Level K4** Undertakes some athletic activities in addition to daily tasks	
	No of AHP Respondents		*No of AHP Respondents*		*No of AHP Respondents*	
	41	*35*	*40*	*34*	*38*	*32*
	Primary	**Established**	**Primary**	**Established**	**Primary**	**Established**
Balance & Coordination	2	2	2	2	3	3
Energy, Drive & Motivation	3	3	4	4	4	2
General Muscle strength	4	4	3	3	2	4
General Range of Movement	7	7	6	6	6	5
Manual Dexterity	10	10	10	9	9	9
Memory	6	8	8	8	8	8
Pain Control	5	5	5	6	5	7
Self Confidence	9	9	7	5	7	6
Socket fit & comfort	1	1	1	1	1	1
Support of family & friends	8	6	9	10	10	9
1 is the most important and 10 is the least important						

The sample of pwLLAs also considered the same factors and their average rankings are listed in table 3.10. The numbers again are represented when 1 is considered the most important factor and 10 the least important.

Table 3.10 Average ranking of importance by pwLLA respondents

Factor	**Rank**
Balance & Coordination	2
Energy, Drive & Motivation	3
General Muscle strength	8
General Range of Movement	5
Manual Dexterity	9
Memory	10
Pain Control	6
Self Confidence	4
Socket fit & comfort	1
Support of family & friends	7

The activity levels of the amputee respondents were considered K3 and above and so it was possible to compare their responses with those of the AHP's by K level.

The average rankings for all the AHP's responses are broadly similar compared to those of the pwLLAs, except for General Muscle Strength and Self Confidence as illustrated in figure 3.1 below. This similarity is also seen when looking at the responses from AHP's for K3 and K4 only, also seen in the scatter plot

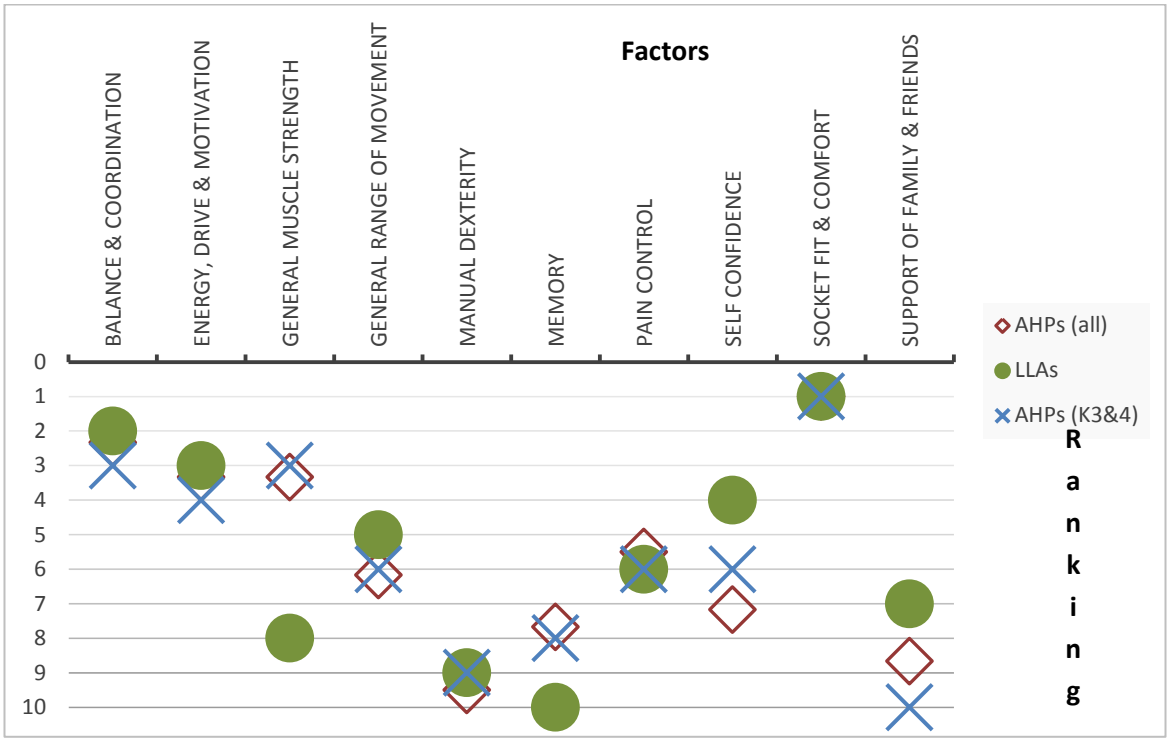


Figure 3-1 Average ranking of factors by AHP's and pwLLA respondents

3.5 Discussion

3.5.1 Respondents and response rates

The survey respondents included a high number of physiotherapists in comparison to the prosthetists and occupational therapists. Regarding the experience and ratio of time all respondents spent working with amputees, most of the responses appeared to have come from a group of AHP specialists. There are 43 Prosthetic Centres across the UK, each employing a MDT to assess and deliver specialist prosthetic rehabilitation. Assuming that each of the physiotherapy respondents represented a prosthetic centre, then this finding may be sufficient to make assumptions for physiotherapists who work with pwLLAs regarding the use of outcome measures. However, there is a need to be cautious when considering the generalisability of these finding for both prosthetists and occupational therapists. The method of distribution to the occupational therapists did not involve any

specialist professional networks and there were much smaller numbers of respondents from them and the prosthetists.

The use of standardised outcome measures is recommended widely (Skinner and Turner-Stokes 2006, Hammond 2000, Jette et al. 2009, Corr and Siddons 2005, Abrams et al. 2006); and potential respondents were asked to return the survey, even if they did not use any outcome measures. However, it could be ventured that those AHP's who do not regularly use outcome measures in their clinical practice, did not wish to complete the survey, even though it was anonymous.

The lack of response by prosthetists and occupational therapists may also be due to the method of distribution, which was through personal contacts. The physiotherapy BACPAR and SPARG members were asked to forward copies to their AHP colleagues in their MDT. It is not always the case that there is an occupational therapist in every prosthetic MDT, therefore the low response rate from them may be due to their scarcity. In contrast, there is a dedicated prosthetist in every prosthetic MDT, so it could be suggested that the lack of response from this profession may indicate that outcome measures are not used regularly in clinical practice. However, only two respondents (both prosthetists) stated that they did not use any, indicating that 82% of prosthetists who completed the survey did use outcome measures regularly. This is in contrast to the findings of a recent survey distributed by BAPO, where it was reported that only 24% of respondents (both prosthetists and orthotists) regularly use outcome measures, with the main barriers being time and insufficient training in the use of outcome measures (Young et al. 2015). The survey response rate for those working in lower limb prosthetics was only 21%. The use of outcome measures among prosthetists in the United States also appears limited and confidence in administering them is low (Gaunaurd et al. 2014).

The use of outcome measures by occupational therapists within the field of lower limb prosthetic rehabilitation has not been recorded, according to the author's knowledge.

3.5.2 Outcome measures

The findings from this survey show that there is a wide range of outcome measures being used by AHP's, during lower limb prosthetic rehabilitation in the UK. The top five outcome measures selected most often were: the SIGAM, LCI / LCI-5, TUG, TWT and the SCS. The original LCI and modified version, LCI-5, were considered together in the current survey. All five were used both by physiotherapists and prosthetists, except the LCI / LCI-5 which was not used by the prosthetists. None of these top five outcome measures were used by the occupational therapists who took part in the survey.

While it is recognized that rehabilitation following amputation should consider both physical and psycho-social factors, all of the top five outcome measures, most regularly used by physiotherapists and prosthetists, measure constructs of physical function or are closely connected, i.e. the interface between residuum and the prosthesis or socket fit in the SCS. Outcome measures concerned with measuring psycho-social factors, such as quality of life or body image, were selected less often by the respondents of this survey.

The LCI / LCI-5, TUG and TWT feature in the BACPAR Toolbox (Cole et al. 2014), the SIGAM Mobility Grades is recommended for use by the BSRM (Skinner and Turner-Stokes 2006), and BAPO recommend the SCS as well as the TUG (Young et al. 2015). It is interesting to note that some respondents reported that they discontinued using the SIGAM and LCI / LCI-5. Reasons cited were: a lack of relevance to clinical practice, an apparent lack of sensitivity to the amputee's changing condition, and a possible ceiling effect when used with higher functioning pwLLAs. Franchignoni et al (2007) reported on the development of the LCI-5, which extended the response categories for each question from four to five, to distinguish whether the amputee used walking aids or not when they scored themselves able to complete a task alone (Franchignoni et al. 2007b). This increased the measure's ability to discriminate between patients and reduced the ceiling effects seen with the original version, especially when it was used with high-functioning amputees. It is possible that some respondents to the survey were commenting on the use of the original version (LCI) with patients at K level 3 or 4, and therefore seeing the ceiling effect described by Franchignoni et al (2007). In contrast to some comments seen in the survey, the SIGAM was reported to be easy to use and not time-consuming by

Rommers et al (2008), who reported on its translation into Dutch (Rommers et al. 2008).

In addition to stating that a lack of time restricted the use of outcome measures, respondents also specifically stated, that if an outcome measure took too long to administer then they would stop using it. For a busy clinician, it is vital that time spent collecting data provides results from outcome measures with proven psychometric properties to ensure correct interpretation.

There is published evidence for the validity and reliability for each of the top five OMs in the survey; TUG (Schoppen et al. 1999, Deathe and Miller 2005, Resnik and Borgia 2011), TWT, both the two minute (2MWT (Resnik and Borgia 2011, Brooks et al. 2001, Brooks et al. 2002)) and six minute (6MWT (Resnik and Borgia 2011, Lin and Bose 2008)), LCI / LCI-5 (Gauthier-Gagnon et al. 1998, Franchignoni et al. 2007b, Miller et al. 2001, Franchignoni et al. 2004), SCS (Hanspal et al. 2003) and SIGAM Mobility Grades (Ryall et al. 2003b). The psychometric properties of LCI / LCI-5 have been assessed most often, with internal consistency reported in four different studies and construct validity reported in three. While reliability has been reported for all the top five, measurement error values have only been published for the TUG, 2MWT and 6MWT.

The one study that reported the reliability and criterion-validity for the SCS, provided data on its responsiveness to socket changes as well as. Responsiveness is also reported for the SIGAM but it is from the same study that reported on its reliability and validity (Ryall et al. 2003). Chapter 4 presents results from the systematic review of the literature that investigates the psychometric measurement properties of outcome measures of physical function used with adults with a lower limb amputation and these studies will be discussed in detail there. Later chapters will also present details and further discussion on the reliability and measurement error (chapter 5) and responsiveness (chapter 6) values for all the top 5 outcome measures.

3.5.3 Factors influencing prosthetic outcome

The pwLLA respondents reflected on their own prosthetic outcome, which was at least at the level of K3 i.e. a “community ambulatory”, when considering the most

influential factors. The AHP respondents considered different outcomes according to the activities level in each K-level. Results showed that the amputee respondents identified “Socket Fit & Comfort”, “Balance & Coordination” and “Energy Drive & Enthusiasm” as the three most important factors that influenced the outcome of their prosthetic rehabilitation. While, the same three factors were reported by the AHPs in most of the patient groups, they also considered “General Muscle Strength” important for primary amputees at the two highest functioning levels (K level 3 and 4).

Three of the top factors identified were similar to two themes of interest to pwLLAs, reported by Legro et al (1999) (Legro et al. 1999). The comparable themes were: “*socket fit*” which relates to the factor “Socket Fit and Comfort” in this survey; and “*other non-mechanical issues*” which relates to “Balance and Coordination” and “General Muscle Strength”. The other two themes mentioned by Legro et al (1999) were: “*adaptation to life and the support of others*“ which could relate to the factor “Support of Family and Friends”, however, this was ranked quite low with all respondents; and “*mechanical aspect of the prosthesis”* which does not relate to any of the ten factors in this survey. Mechanical aspects of the prosthesis, beyond the fit of the socket, were not included as understanding the external influence of different prosthetic components could form a single topic for another survey. Although support of family and friends were not highly rated in this survey, a meta-analysis of qualitative literature, conducted sometime after the completion of this survey, did recommend that this group of patients make greater use of patients’ social support networks, peer support from other pwLLAs, as well as professional psychotherapy support (Murray and Forshaw 2013). The aim of the meta-analysis was to provide a more detailed understanding of some of the “common and recurring” areas of importance to adults who had undergone a lower limb amputation and were aiming for successful prosthesis.

Energy levels have previously been shown to relate to quality of life (Demet et al. 2003). It was therefore not surprising that “Energy, Drive and Motivation” was chosen by both the pwLLAs and AHPs as one of their top three factors here. Motivation was also considered one of the most important predictors in the rehabilitation process following a Delphi Study which involved prosthetic users and health professionals (Schaffalitzky et al. 2012).

AHP respondents ranked “General Muscle Strength” of high importance (ranked two) for the more active amputees, though the pwLLA respondents ranked this factor, on average, the eighth most important. It may be that they took this factor for granted, as it is likely that their muscle strength would have improved over the course of their therapy which is planned by the AHP’s. For the pwLLAs, “Self-confidence” was much more important and was listed at number four in their ranking, whereas AHP’s did not rank it higher than five, and mostly much lower. Self-confidence or self-efficacy is known to affect performance (Bandura 2010, Ryan and Deci 2000), so it is important to take note of the influence that self-confidence and self-efficacy can have on physical performance.

3.5.3.1 Do the most widely used outcome measures measure constructs that align to the most influential factors?

It is easy to see why “Socket Fit and Comfort” is critical to successful prosthetic rehabilitation and the SCS is used to quantify how good or bad the fit is, as perceived by the comfort of the socket. If the interface between the amputee and prosthesis does not fit and is uncomfortable, all activities using the prosthesis will be compromised.

The remaining top four outcome measures all measure aspects of mobility, quantifying how well the prosthetic outcome has been achieved in terms of distance, time and level of independence. However, none of the top five outcome measures assessed the factors: balance and coordination, energy, drive and enthusiasm, or general muscle strength directly.

There is a high correlation between the Activities-specific Balance Confidence (ABC) Scale-UK, the 2MWT and the TUG (Miller et al. 2003) and therefore improvements in balance are likely to show improvements in the times (TUG) and distances (2MWT) recorded and vice versa. However, if good balance and coordination are considered important in achieving a successful prosthetic outcome by the AHP’s, it can be argued that measures focused on balance, e.g. the ABC and BBS will be more successful in establishing the amputee’s balance capability. Results from these specific balance scales can help focus therapy programmes, to facilitate better balance and subsequently better prosthetic use for the pwLLA. While there were some AHP’s who reported using the ABC Scale-UK, many more

reported using a TWT and / or the TUG. This may have been because they are easier to implement, i.e. quicker to use, simple to understand and require less equipment, or it could be because the AHP's are just more familiar with a TWT and the TUG. The BBS is a familiar outcome measure to many physiotherapists working with the elderly, but it was not included in the survey. This was because, at the time of designing the survey, there was no published evidence of its validity and reliability with pwLLAs. A study by Major et al (2013) has since shown the BBS to have high inter-rater reliability and good internal consistency in pwLLAs (Major et al. 2013, Wong 2013, Wong et al. 2013). Wong et al (2013) also reports good construct validity for the BBS, but there was a suggestion of a ceiling effect with some high functioning amputees scoring at the top of the scale (Wong 2013). Interestingly, none of the respondents of this survey reported using the BBS with pwLLAs.

Energy, drive and enthusiasm are difficult concepts to measure as they are largely subjective, but hugely important. With the link between emotional well-being and physical activity well documented (Bherer and Liu-Ambrose 2013), it could be argued that if a pwLLA has more energy their physical performance overall is likely to improve. Determination to walk and motivation were considered important predictors in the rehabilitation process by prosthetic users and health professionals (Schaffalitzky et al. 2012). Patient-reported outcome measures, designed to capture the patient's experiences and perspective, and especially those that measure Quality of Life may be more suited to interpreting how much drive and enthusiasm the person has. However, these types of outcome measure were not favoured by the AHP's in this survey. It is not known why, though one possible explanation may be that physiotherapists, the largest number of respondents, use outcome measures primarily to assess physical function. Physical function is the primary aim of physiotherapeutic interventions, and although the impact of psychosocial factors on the amputee's physical performance is understood, they may not feel it necessary to overtly measure these factors. This could also be true for prosthetists, but is likely to be less so for occupational therapists as they often aim to address psychosocial factors in their assessments and treatments.

Respondents acknowledged the importance of both physical and psycho-social factors in a successful prosthetic outcome. Therefore the significance of measuring

both the amputee's physical and psychosocial state following any intervention should be emphasised.

3.6 Limitations

With the relatively small number of respondents in this survey, (45 of which 23 were physiotherapists) the generalisability of the results to AHP's working with pwLLA clinical practice in the UK is limited.

While the response rate was low, it was encouraging to see the experience of the respondents with amputee rehabilitation. The aim of this survey was to gain information from those AHP's who were working with amputees during prosthetic rehabilitation. This aim was achieved with the majority of respondents having worked with amputees for at least 5 years and also working with this population for most of their working week. Circulating the survey through Professional Special Interest Groups working with amputees helped achieved this.

The method of distribution especially for the prosthetist and occupational therapist group could have been improved by following up initial e-mails to secretarial staff of the networks.

The limited number of pwLLA respondents was disappointing and therefore further exploration of the factors considered important to them following prosthetic rehabilitation is required to further confirm or refute the findings of this small sample group.

The choice of factors was limited to ten. With only one additional factor added by any of the respondents it could be considered a comprehensive choice. However, with the small sample size, this is another aspect that could be explored further to confirm the inclusiveness of the list.

3.7 Conclusions

AHP's use a wide variety of outcome measures during prosthetic rehabilitation in the UK, though not all professions use them regularly. This variety of outcome measures used indicates there is still a lack of consensus on which outcome measures to use. The group of outcome measures most commonly used, all measure constructs in the physical domain.

The ranking of the most important factors influencing a successful prosthetic outcome was found to be similar in the pwLLAs and AHP respondents, with socket fit and comfort rated the most important by both groups. The factors considered important by both groups are not necessarily reflected in the choice of outcome measures that had been used regularly by the AHPs, except for the SCS. However the outcome measures did provide a measure of the quality of the prosthetic outcome.

Although the sample size was small, the specialist nature of the AHP respondents gives authenticity to the results within the context of prosthetic rehabilitation.

The results from this very small sample of pwLLAs must be considered with caution and it is recommended that further work is carried out to confirm the findings from this group.

4 Chapter 4 Systematic Review (study II)

4.1 Purpose of chapter

The purpose of this chapter is to describe the methodology, and to present and discuss the results, of a systematic review of the literature that investigated the psychometric measurement properties of outcome measures of physical function used within populations of pwLLAs (adults only). Results for the assessment of the methodological quality and strength of evidence presented in the studies will be reported, and implications for clinical practice and future research will be discussed.

The systematic review was undertaken to answer the following research questions:

Research question 2

Which outcome measures, used to measure physical function during prosthetic rehabilitation, have been investigated and published in peer-reviewed journals; and what are the characteristics (i.e. psychometric properties) that have been presented?

Research question 3

What is the methodological quality of the published studies presenting results for psychometric measurement properties of the outcome measures used to measure physical function during prosthetic rehabilitation?

Research question 4

What are the levels of evidence presented in these studies?

The extent of the amputees' capacity for physical function (and other related aspects of capability) that might have been assessed by outcome measures during prosthetic rehabilitation will not be considered in this chapter. Rather, the quality and how strong the evidence is for judgement on that question will be the focus for this systematic review.

4.2 Background

Financial accountability within the area of prosthetic rehabilitation, has led to a renewed interest in the ability of outcome measures to measure the efficacy of many interventions along the amputees' care-pathway. For example: "What is the effect of new surgical procedures" (Sullivan et al. 2003, Frossard et al. 2010); "which prosthetic componentry provides the most benefit" (Kannenberg et al. 2014, Gailey et al. 2012); and, "what is the long-term functional performance of our patients after discharge from rehabilitation" (van Twillert et al. 2013).

Research has shown that the proper selection of appropriate outcome measures is crucial for clinical effectiveness (Johnston and Dijkers 2012). The correct choice of outcome measure is critical to understanding the status of, or demonstrating change in, any patient's condition, but correct interpretation of the results is also essential. A full understanding of the psychometric measurement properties of all outcome measure will permit the selection of the most appropriate one(s), in the right context, ensuring that the results are interpreted appropriately.

Of the many constructs that may influence the outcome of prosthetic rehabilitation, consideration of physical function following lower limb amputation surgery, is a basic requirement (van Velzen et al. 2006). Many studies have reported the development, and subsequent use, of outcome measures that measure constructs related to physical function of the pwLLA. These studies have been performed in a variety of settings; before, during and after limb-fitting. As highlighted in chapter 2 previous reviews have reported on the measurement properties of such outcome measures, though to varying degrees. Rommers et al detailed 35 different outcome measures that measured mobility in a review published in 2001. There was little discussion of the psychometric properties of the measures, beyond commenting that test re-test reliability and validity were carried out. Reliability was considered in three categories (test retest, inter-rater and internal consistency), with validity considered in four categories (face, content, construct and criterion) for the outcome measures reviewed by Condie et al in 2006. However, only the presence or absence of appropriate values in each category were recorded for each study, using a checklist similar to that published by Jerosch-Herald in 2005 (Jerosch-Herold 2005). The authors also admitted that time did not allow them to undertake a comprehensive methodological quality review of the studies (Condie et al. 2006).

Three reviews (Deathe et al. 2009, Hebert et al. 2009, Hawkins et al. 2014) used guidelines published by Johnston & Graves (2008). These criteria outline aspects of reliability and validity that should be presented when developing or testing a measure in a new population, with sensitivity to change considered within validity. The absence, or presence, of these different aspects of reliability and validity are recorded, and an overall quality rating is given. The authors acknowledged that a multilevel grading of components of measurement quality would allow a grading to be given for each measurement property and "deserves investigation" (Johnston and Graves 2008).

The psychometric measurement properties of outcome measures that measure physical function of pwLLAs have been previously presented. However, the level of the evidence has not been presented per measurement property for each outcome measure, taking into account the methodological quality rating for the individual studies presenting the evidence.

The aims of this systematic review are therefore, to: examine the published literature investigating outcome measures used to measure physical function of pwLLAs; provide a methodological quality-rating for each psychometric measurement-property presented; and report the level of evidence obtained for each outcome measure. A synthesis of the methodological quality and level of the evidence will be provided for each outcome measure to enhance the understanding of their utility. This will allow both the clinician and researcher to make a more informed choice about which outcome measure(s) to use from the many available.

4.3 Methods

4.3.1 Study Selection and Data Collection Process

4.3.1.1 Literature Search Strategy

The Cinahl, Medline, PsycInfo, SCOPUS and Pubmed databases were searched in December 2013 using the following search terms:

[lower limb amp* OR lower-limb amp* OR LLA OR transtibial OR trans-tibial OR transfemoral OR trans-femoral]
AND [outcome measure* OR test OR score OR scale OR assessment OR questionnaire OR instrument OR index OR tool]
AND [clinimetric properties OR validity OR reliability OR responsiveness OR repeatability OR sensitivity OR internal consistency]

As wide variety of search terms can be used when referring to psychometric measurement properties it was decided to incorporate a specific search strategy that had been developed, to check that no titles had been missed. Two search filters for finding studies on measurement properties in Pubmed was published in 2009, (Terwee et al. 2009). One search filter is highly sensitive for finding studies on measurement properties with a sensitivity of 97.4% and a precision of 4.4%. While the other is a more precise search filter which needs less abstracts to be screened, but is at a higher risk of missing relevant studies. This filter has a sensitivity of 93.1% and a precision of 9.4%. Both filters are designed to be used in conjunction with specific construct and population searches, as required.

The sensitive search filter described by Terwee et al (2009), was used together with the population search terms described above to search Pubmed. Versions of the filters are now available for other databases on the COSMIN website, but were not at the time of the initial search. See Appendix 4 for full details of the search terms.

Filters were set in all databases to include adult participants and English text only. Dissertations, conference proceedings, editorials, opinion pieces, review studies, letters, single case studies and case series of four or fewer patients were excluded.

Studies were also excluded if they: investigated outcome measures used to predict future outcome e.g. length of hospital stay or prosthetic usage; reported outcome measures used exclusively for measuring quality of life; examined the performance of either prosthetic componentry or instrumentation rather than that of the pwLLA; or were validation studies of translated versions of any outcome.

Titles resulting from the searches were initially reviewed by one reviewer (PhD student, JS) to exclude any duplications and obviously irrelevant studies. The same

reviewer then categorized all studies as “include”, “exclude” or “possible” with reference to the following PICO criteria:

Population:	Lower limb amputees, unilateral or bilateral at any level
Intervention:	Assessing measurement properties of outcome measures of physical functional
Condition:	Any stage of prosthetic rehabilitation
Outcome:	Reporting psychometric measurement properties

Reference lists of included studies were also scrutinised and any relevant studies were added. A second reviewer (PhD Director of Studies, MvdL) then checked the titles against the inclusion and exclusion criteria. If there was any doubt or disagreement over relevance and / or adherence to the criteria, abstracts were read by both reviewers. If there was any disagreement or doubt about including the paper at this stage it was submitted for full text review by both reviewers. The reasons for all exclusions following abstract and full study review were logged and discussion notes between the two reviewers were logged throughout the process to compile “Reviewers’ notes“- see Appendix 5. This helped to ensure consistency between the reviewers.

4.3.2 Methodological quality analysis

The methodological quality of studies selected for inclusion in the review was independently assessed by the same two reviewers. This was undertaken using the COSMIN checklist (Terwee et al. 2012). The COSMIN checklist was chosen as the quality review tool because of the multilevel grading of each measurement property within a standardised rating system and the extensive guidelines that accompany the checklist (Mokkink et al. 2012). Standardisation of the reporting of systematic reviews is recommended by many (Moher et al. 2007, Moher et al. 2009, Gianola et al. 2013, Shamseer et al. 2015, van Tulder et al. 2003) and COSMIN provides detailed standards specifically for the preferred design characteristics and statistical methods of studies assessing measurement properties (Terwee et al. 2012). The COSMIN standards are similar, but distinct, from those recommended for assessing studies measuring the effects of healthcare interventions (Furlan et al. 2015). They expand on the requirements for individual measurement properties within previous

checklists (Johnston and Dijkers 2012) and the COSMIN checklist gives a grading for each measurement property, not the outcome measure as whole.

The COSMIN checklist and the accompanying four-point scoring system of excellent, good, fair and poor (see Appendix 6) provides an overall score for the methodological quality of a study for each category of measurement property being examined. Taxonomy of terms and their definitions, also provided by the COSMIN group (Mokkink et al. 2010b), facilitates agreement between reviewers when different terminology is presented across studies. There are nine categories outlined in the COSMIN checklist, each with a different number of items to be considered (Mokkink et al. 2010c). These include: sample size; management of missing data; choice of statistical analysis; choice of gold standard or comparator instrument; and conduct of the study in terms of implementation and administration of the data. In the Internal consistency category, there are 11 items, for Reliability (14 items), Measurement error (11 items), Content validity (5 items) Structural validity (7 items), Hypotheses testing (10 items) Cross-cultural validity (15 items), Criterion validity (7 items), Responsiveness (18 items) and Interpretability (9 items). All but one (Cross-cultural validity) of these categories were considered in this systematic review. Comments on the generalisability of the results with regard to the population under study and the setting, were also to be included, but this section is not scored in the COSMIN checklist.

An Excel™ spreadsheet was used to collect data in each of the relevant categories against the COSMIN 4-point scale of excellent, good, fair and poor. The overall rating of each category was given as the lowest level response to any item within that category, as per the COSMIN guidelines.

Where applicable, the rating for sample size within the methodology review was: excellent for n>100, good n=50-99, fair n=30-49 and poor n<30. However, sample size was not considered during the methodological quality review. This was to avoid penalising studies with small sample sizes twice.
Though not explicitly stated in the guidelines, an overall quality rating of each study was also calculated using the same “worst score counts” principle, across all the measurement properties presented in the study.

4.3.3 Levels of evidence analysis

The strength of the results provided within each study was also examined using published criteria (Terwee et al. 2007) (see Table 4.1). The three possible levels of evidence are given in this checklist. They are positive "+", indeterminate "?", and negative "-". Where no information is available a "0" rating is given.

Table 4.1 Levels of evidence - taken from (Terwee et al. 2007)

	+	?	-
Internal consistency	Factor analyses performed on adequate sample size (7 x # items and ≥100) AND Cronbach's alpha(s) calculated per dimension AND Cronbach's alpha(s) between 0.70 and 0.95	No factor analysis OR doubtful design or method	Cronbach's alpha(s) <70 or >0.95, despite adequate design and method.
Reliability	ICC or weighted Kappa ≥0.70	Doubtful design or method (e.g., time interval not mentioned)	ICC or weighted Kappa <0.70, despite adequate design and method
Measurement error	Minimal Important Change (MIC) < SDC OR MIC outside the LoA OR convincing arguments that agreement is acceptable	Doubtful design or method OR (MIC not defined AND no convincing arguments that agreement is acceptable)	MIC ≥ SDC OR MIC equals or inside LoA, despite adequate design and method
Construct Validity	Specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses	Doubtful design or method (e.g., no hypotheses)	Less than 75% of hypotheses were confirmed, despite adequate design and methods
Responsiveness	SDC or SDC < MIC OR MIC outside the LoA OR RR >1.96 OR AUC ≥ 0.70	Doubtful design or method	SDC or SDC ≥ MIC OR MIC equal or inside LoA OR RR ≤1.96 OR AUC >0.70, despite adequate design and methods

For studies with a sample size of < 50 participants, an indeterminate "?" rating is given in several categories as the authors of the criteria determined this to be the number adequate for the statistical analyses (Terwee et al. 2007). In this review, sample size was considered when reporting the levels of the evidence and not when reporting on the methodological quality. This procedure had also been adopted in

many previous rehabilitation systematic reviews involving, for example: osteoarthritic hip and knee populations (Dobson et al. 2012), gait related outcomes in children with cerebral palsy (Ammann-Reiffer et al. 2014) testing 6MWT in children (Bartels et al. 2013) and testing 2MWT in adults (Pin 2014).

4.3.4 Best evidence' synthesis

Methodological quality ratings for the measurement properties of each outcome measure were presented alongside the level of evidence in each study. This allowed inferences to be made on the relative robustness of evidence for each outcome measure.

A third reviewer was available, though not used, to resolve any disagreements between the reviewers over the level of ratings or methodological quality, or during the examination of the results and levels of evidence.

4.4 Results

4.4.1 Search results

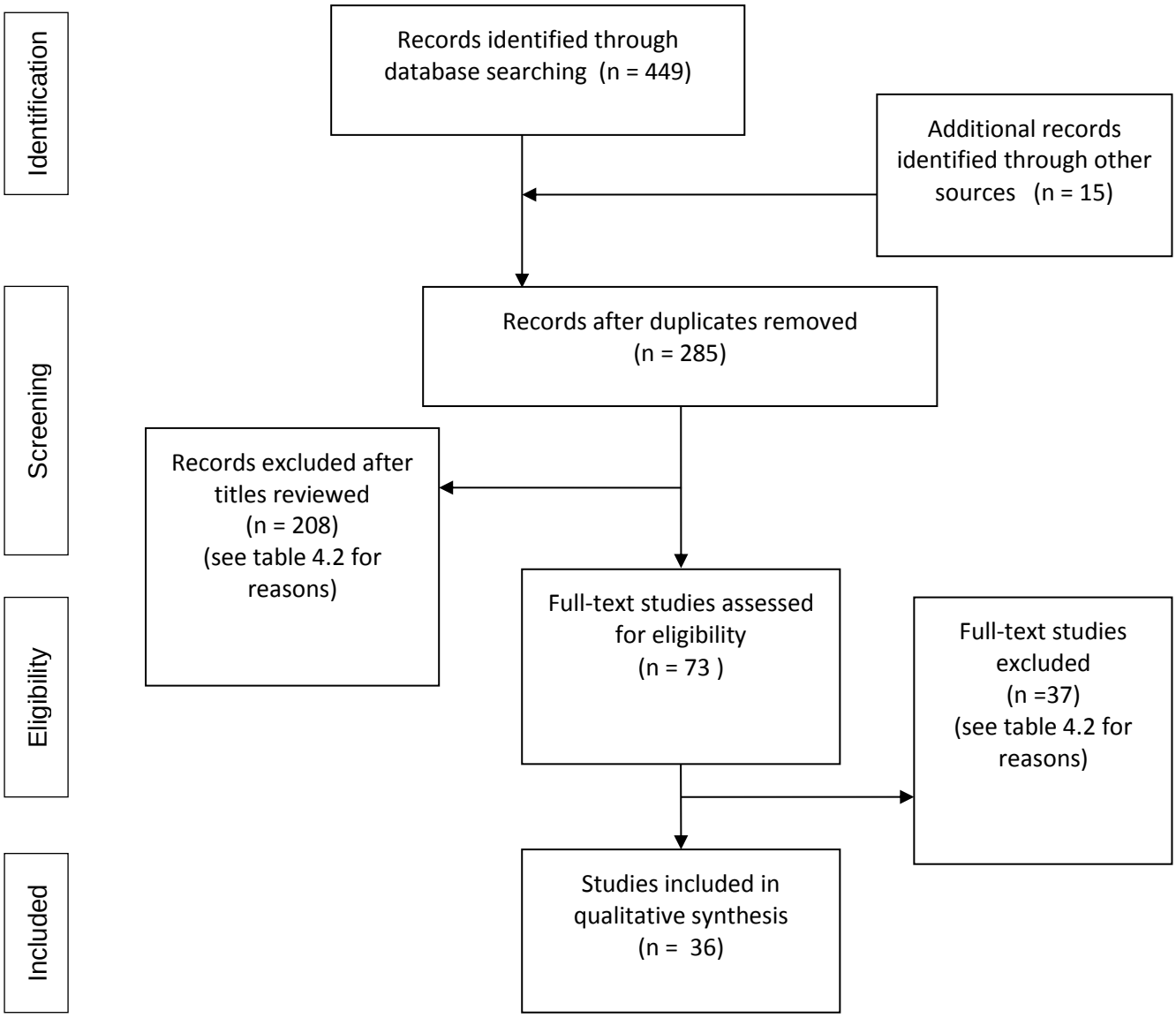


Figure 4-1 Search strategy results

A total of 449 applicable titles were retrieved from all databases using the basic search terms: 222 from Cinahl, Medline, and PsycInfo, 219 from Pubmed and 8 from SCOPUS. Additional sources, including reference lists of related studies, were also scrutinized to identify studies not found through the electronic searches and this produced a total of 15 additional titles. The relatively high number of additional titles did raise some concerns about the sensitivity of the search terms. This will be discussed later in the limitations section of the chapter. When duplicates were removed a total of 261 studies remained see Figure 4.1.

After reviewing the titles and abstracts against the inclusion / exclusion criteria, 73 were retrieved for full text scrutiny. No additional titles were added following the search using the published search filter terms, by Terwee et al (2009). Inspection of the full text, and a further examination of the inclusion/exclusion criteria, resulted in the reviewers agreeing that 36 studies were to be quality assessed against the COSMIN scoring system. Table 4.2 shows a breakdown of the reasons why studies were rejected firstly on the basis of the title and then on the basis of the abstract or full text review.

Table 4.2 Reasons for excluding studies

Reason for rejection	**Number rejected on basis of title**	**Number rejected after reading full text**
Not relevant to main aim	197	3
Main focus was translation	7	2
Main focus was Quality of Life	2	5
Review papers i.e. no new data	2	2
Sample size too small or a paediatric population	0	2
Main focus was measurement of instrumentation not performance of subjects	0	11
No psychometric data presented	0	9
Main focus was predictor tools	0	3
TOTAL	208	37

4.4.2 Descriptive results

The final 36 studies in the review investigated evidence from a total of 2,688 adult pwLLAs. Sample sizes ranged from 8 to 448 and measurement properties were reported for 37 outcome measures. Findings for two outcome measures were excluded from the results because they did not fit the inclusion criteria; the Amputee Mobility Predictor is used predominantly for predicting outcomes and the SF-36 is a measure of quality of life. However, the study reporting on these two outcome measures was included in the review because it presented data on the measurement properties of six other outcome measures (Resnik et al 2011).
Table 4.3 presents the number of studies included in the review and which measurement properties that were reported on. See Chapter 2 for a discussion about the variety of terms used to describe similar properties and their full definitions. The number of outcome measures on which results were presented on is also presented.

Table 4.3 Number of studies with number of outcome measures presented per measurement property reported

Measurement Property	Studies	Outcome Measures
Internal Consistency	16	9
Reliability	27	29
Measurement Error	15	18
Validity	22	20
Responsiveness	8	11
Total	36	35

While the 36 studies provided sufficient data to populate all measurement-property categories, not all outcome measures received full psychometric scrutiny. Overall, a total of 27 studies reported reliability parameters for 29 outcome measures. Some studies presented a single aspect e.g. inter-, intra-rater or test re-test reliability, while others studies presented more than one. Content validity was not formally reported by any of the studies, though some comment was made where factor analysis had been performed. It is agreed that “gold-standard” criteria are not yet established for this population, and thus were unavailable for use by any of the studies. Nevertheless, construct validity of outcome measures was reported in

some of the studies by testing hypotheses stated before analysis of the data (*a priori*).

Details of the aetiology and levels of amputation that the participants had undergone are given in Table 4.4. This table also outlines the range of study designs and methodologies that were used. Most studies included all aetiologies in their samples, though some studies did recruit only vascular participants (Schoppen et al. 1999, Panesar et al. 2001) or only non-vascular participants (Hagberg et al. 2011). The study participants presented with a variety of amputation levels but the majority were trans-tibial and / or trans-femoral.

Table 4.4 Demographic and study-design details for all included studies

Author(s) Date	OMs studied	Population		Study Aim	Study Design	Sample size
		Amp Level(s)	Aetiology			
Brooks et al 2001	2MWT	Mixed levels of amputation, both uni and bilateral	Mixed	Determine **construct validity** and **responsiveness** of 2MWT	Data was extracted from a rehabilitation database. The 2MWT times and SF-46 scores were obtained twice during admission: following initial fitting and 48hrs prior to discharge, and at 3mths post-discharge. Houghton Scale scores only obtained at discharge.	Convenience sample n=290 Construct Validity n=56 (Houghton) to 142 (SF-36) Responsiveness n=197
Brooks et al 2002	2MWT	Unilateral TTAs	Mixed	Determine **inter** and **intra-reliability** of the 2MWT	Participants, both in-patients and out-patients, completed two successive walks measured by two different raters on two consecutive days. The order of the raters was reversed on the second day.	n=33
Callaghan et al 2002	The Functional Measure for Amputees (FMA)	Unilateral TTA,	Mixed	Describe development of a new outcome measure to collect long-term functional data, and determine its **test-retest reliability**.	The FMA was developed from the Prosthetic Profile of the Amputee (PPA). Two FMA questionnaires were posted to participants 4 weeks apart to assess reliability.	n=133

de Laat et al 2010	Climbing stair Questionnaire	Mixed levels of amputation, both uni and bilateral	Mixed	Determine **construct validity** and **test-tetest reliability** of the Climbing stair Questionnaire	Participants were recruited during a 10 year period completed the questionnaire via a postal survey. Those repeating the survey 3 weeks later were asked to confirm their status was stable before completing the 2nd questionnaire.	Construct Validity n=172 Reliability n=24
de Laat et al 2011	Rising and sitting down Questionnaire	As for De Laat et al 2010	Mixed	Determine **construct validity** and **test-retest reliability** of the Rising and sitting down Questionnaire	As for De Laat et al 2010	Construct Validity n=171 Reliability n=22
de Laat et al 2012	Walking Questionnaire	As for De Laat et al 2010	Mixed	Determine the **construct validity** and **test-retest reliability** of the Walking Questionnaire.	As for De Laat et al 2010	Construct Validity n= 172 Reliability n=22
Deathe et al 2005	L-test	Unilateral TTAs & TFAs	Mixed	Assess the **concurrent validity** and **intra and inter-reliability** of the L-test	A consecutive sample of participants were asked to complete a series of walk tests and a series of self-report questionnaires. A second rater conducted a second series of tests later in the same visit to assess inter-rater reliability. Re-testing of all walking tests and questionnaires was undertaken 2 weeks later.	Concurrent Validity n=93 Reliability n=27

Devlin et al 2004	Houghton Scale	Mixed levels of amputation, both uni and bilateral	Mixed	Evaluate **responsiveness** and ceiling and floor effects of the Houghton Scale. Confirm previous **test-retest reliability** and **convergent validity**	Two convenience sample groups were recruited. Sample 1 - Scores were collected, first by telephone interview and then again a week later. Sample 2 - Data were collected at two time points, three months apart	Test-retest reliability n=49 Responsiveness, & validity n=76
Franchignoni et al 2003	Rivermead Mobility Index (RMI)	TTA and TFA	Mixed	Examine **internal consistency**, **validity**, **responsiveness** and scalability of the Rivermead Mobility Index	A convenience sample of patients were recruited and data was collected at admission and discharge. In addition the 10m walk test was performed as soon as the patient progressed out of the parallel bars.	Internal consistency & responsiveness n=140 Construct validity n=70
Franchignoni et al 2004	Locomotor Capability Index (LCI) and revised LCI-5	Unilateral lower limb amputation	Mixed	Assess **reliability**, **validity** and **responsiveness** of both versions of LCI	Data was collected from consecutive patients undergoing prosthetic training, on admission and at discharge.	IC, construct validity & responsiveness n= 50 Test-retest n=37
Franchignoni et al 2007a	Prosthesis Evaluation Questionnaire (PEQ Mobility Section (ms))	Mixed levels of amputation, both uni & bilateral	Mixed	Perform a **Rasch analysis** of the PEQ (MS) and confirm **validity** and **reliability**	Consecutive sample who had completed a prosthetic rehab programme were sent a questionnaire pack.	n=123

Franchignoni et al 2007b	LCI-5	Mixed levels of amputation, both uni and bilateral	Mixed	Perform a **Rasch analysis** of the LCI-5 and confirm **validity** and **reliability**	Consecutive sample who had completed a prosthetic rehabilitation programme were sent a questionnaire pack.	n=123
Gardiner et al 2002	Harold Wood - Stanmore Mobility Scale, Handicap Mobility Scale, Handicap Physical Independence Scale.	No demographic detail was available for the patients included in the study.		Assess **inter-rater reliability** for three disability/handicap scales.	Four assessors completed admission and discharge scores on pwLLAs during in-patent prosthetic rehabilitation.	n=14
Gauthier-Gagnon et al 1994	Prosthetic Profile of the Amputee Questionnaire (PPA)	Unilateral TTAs and TFAs	Mixed	Assess **test-retest reliability** and **construct validity** of the PPA.	The PPA was sent on two occasions to pwLLAs after they had been discharged from hospital.	n=60 (sample that completed questionnaires on both occasions)
Gauthier-Gagnon et al 1998	LCI	Mixed levels	Mixed	**Internal consistency** and report on **content validity** of LCI	PPA questionnaire (which contains the LCI was sent to 89 adult pwLLAs who had completed their prosthetic rehabilitation.	Response rate 81% n=70

Hagberg et al 2010	Physiological Cost Index (PCI)	Unilateral mixed levels	Non-vascular only	Evaluate **reliability** of the PCI	pwLLAs with at least three months prosthetic usage underwent within-day test-retesting, of continuous walking at comfortable walking speed, by the same rater. HR was logged at each minute and perceived excertion was recorded using the BORG scale.	n=28
Highsmith et al 2013	Hill Assessment Index (HAI)	Unilateral TFAs	Mixed	Evaluate the **inter-rater reliability** of the HAI	Ssubjects were asked to walk down a 0.9m wide and 4.9m long ramp at an angle of 5^{o}, both before and 90 days after switching from a non-microprocessor knee to a C-leg. The trial was timed, video-taped and independently scored by two raters.	n=21
Hillman et al 2010	Prosthetic Observation Gait Score (POGS)	Unilateral mixed	Mixed	Evaluate the **repeatability** of the POGS.	All participants were videoed walking at self-selected walking speed over a flat surface. Six observers scored these videos on two occasions two weeks apart using POGS.	n=10
Kohler et al 2011	ICF checklist	Mixed	Mixed	Feasibility, of using ICF checklist as an outcome measure **Validity** & **responsiveness**	Checklist used at 4 time-points: pre-adm (self-report recall), 1 week post-amputation (observed), at discharge (observed) and 3mths post-amp (phone call)	n=20

Kristensen et al 2014	One leg stand test (OLST)	Mixed	Non-traumatic amputees	Explore how many one leg standing tests (OLST) is need to provide a stable result, and investigate of proposed protocol.	Participants were recruited early in the rehab period (ave 16 days post-amp). Two tests were carried out on the same day.	Protocol setting n=36 Reliability n=30
Lin & Bose 2008	Six minute walk test (6MWT)	TTAs	Mixed	Investigate **test-retest reliability** and **construct validity** of the 6MWT.	Three trials of 6MWT were conducted in one day and two weeks later the subjects carried out timed up and go (TUG) and one leg balance tests	n=13
Major et al 2013	Berg Balance Scale (BBS)	Mixed	Mixed	Evaluate **validity** and **reliability** of the BBS	The BBS was conducted twice in one day by two independent raters with 20mins rest between each test.	n=30
Miller et al 2004	Frenchay Activities Index (FAI)	Unilateral TFA & TTA	Mixed	Establish **reliability** and **validity** of the FAI	Subjects who had had their prosthesis at least 6 months were recruited if considered stable. They completed the PROMS and walking tests at a clinic appointment. Two weeks later participants either completed the FAI again at the clinic (n=29) or they completed it a home on receipt of a postal copy (n=55).	n=84

Miller et al 2001	The Houghton Scale PPA-LCI, PEQ-ms	Unilateral mixed levels	Mixed	Establish **reliability** and **validity** of the Houghton scale and replicate **reliability** and **validity** for the PPA-LCI and the PEQ - ms	Sample 1 (n=60) were recruited from an OP database of pwLLAs who were stable and walking tests and questionnaires were completed while attending an OP clinic. A second set of questionnaires were sent in the post 4 weeks later to complete. Sample 2 (n=329) were recruited from the same facility but subjects had been discharged for at least 1 year. They received a set of questionnaires in the post on one occasion.	Internal consistency and reliability n=60 Validity n=329
Miller et al 2003	Activities-Specific Balance Confidence (ABC) Scale	Unilateral TT and TTAs	Mixed	Establish the **internal consistency**, **reliability** and **validity** of the ABC Scale	Sample 1 (n=50) were recruited from an OP database of pwLLAs who were stable. Walking tests and questionnaires were completed while attending an OP clinic. A second set of questionnaires were sent in the post 4 weeks later to complete. Sample 2 (n=329) were recruited from the same facility but subjects had been discharged from at least 1 year. They received a set of questionnaires in the post on one occasion.	Internal consistency & reliability n=50 Validity n=329

Monteiro et al 2013	DSF-84 Checklist	16-52 adult male amputees. No details given on levels	No details given	Develop a new instrument devised to assess the functional and social performance. Also to establish **reliability**.	138 subjects participated in finalizing the items included in the checklist and 30 were filmed answering the checklist then four assessors viewed the footage twice to give their scores.	Reliability n=30
Panesar et al 2001	Office of Population Censuses and Surveys (OPCS) Functional Independence Measure (FIM) Amputee Activity Score (AAS)	Mixed levels	Vascular disease	Assess **validity** and **responsiveness** of three established outcome measures used to measure levels of disability in early pwLLA rehab.	Scores were collected on admission, discharge and at 8 weeks following discharge from rehabilitation.	n=34
Resnik et al 2011	PEQ, 2MWT, 6MWT, TUG, OPUS and PSFS,	Mixed levels	Veteran population only	Item analysis of modified PEQ and estimate the **reliability** and **measurement error** for all outcome measures.	Subjects completed all outcome measures on two occasions 1 week apart. The same rater carried out the testing procedures.	n=44`

Rushton et al 2002	Goal Attainment Scaling (GAS)	Unilateral mixed levels	Mixed	Establish **inter-rater reliability**, **validity** and **responsiveness** of the GAS with pwLLAs	A GAS assessment was completed within the first week of a rehabilitation programme by one of two raters with a second assessment taking place 48 hours later. A final GAS assessment took place at discharge.	n=10
Ryall et al 2003b	Special Interest Group in Amputee Medicine (SIGAM) mobility grade	Mixed	Mixed	Develop a protocol and establish **reliability**, **validity** and **responsiveness** for the SIGAM mobility grade	Subjects recruited following completion of their prosthetic rehabilitation. Study 1 (part A) Subjects completed the SIGAM and a timed walk test on two occasions, 2-4 weeks apart - repeatability. (part B) SIGAM is a single-item scale so another measure was conducted at the same time, their results were combined in a "matrix of mobility" and a Rasch analysis was conducted Study 2 investigated the responsiveness of the scale by completing the SIGAM at least twice throughout the rehabilitation period.	Reliability n=62 Validity n=200 Responsiveness n=33
Ryall et al 2003a	Rivermead Mobility Index (RMI)	Mixed	Mixed	Establish **reliability** and **validity** for RMI in pwLLAs.	Subjects were following completion of their prosthetic rehabilitation. Part A: the RMI was independently completed twice on the same patient by two raters on the	Reliability n=62 Validity n=200

					same day. One rater also completed a second assessment two - four weeks later on the same group of patients. Part B: through stratified sampling more pwLLAs were recruited at the first time point to achieve a more representative sample for this part which involved a Rasch analysis to investigate the construct validity of the RMI.	
Sakakibari et al 2011	Activities-Specific Balance Confidence (ABC) Scale	Mixed	Mixed	Evaluate **reliability** of the ABC scale	Survey and chart review data from three previous studies. The data collected previously was grouped into each of the response formats and then entered into the two models chosen (Partial Credit Model and Rating Scale Model).	n= 448
Schoppen et al 1999	TUG	Unilateral TTA or TFAs aged 60 +	Vascular disease only	Establish **inter** and **intra-rater reliability** and **validity** of the Timed Up and Go (TUG) in a population of pwLLAs	Subjects who were deemed stable were recruited and studied in their homes using standardized equipment. Two different raters assessed the subjects twice on the same day. The same rater then assessed the same subject two weeks later.	n=32

Theeven et al 2010	Assessment of Daily Activity Performance in Transfemoral amputees (ADAPT)	Unilateral TFAs.	Majority trauma	Determine the feasibility and test-re-test **reliability** of a new assessment of activity performance measured by ADAPT	Subjects completed a circuit of 9 activity stations and after one hour resting period they repeated the circuit.	n=20
Wong et al 2013	Berg Balance Scale (BBS)	Mixed	Mixed	Assess the **internal consistency** of the BBS and determine its **validity** with pwLLAs who live in the community.	BBS data was collected on one occasion. Four physical therapists and four physical therapy students were trained in the use of the BBS following a standardized protocol.	n=40
Wong et al 2013	BBS	Unilateral; TTA (3), TFA (1) and one ankle dis-articulation	Mixed	Investigate the inter-rater **reliability** of the BBS when used with pwLLAs and the impact of using raters with various levels of clinical experience (registered clinicians and students).	Subjects were video-recorded undertaking the BBS. Two clinicians independently rated the pwLLAs at the time of recording. Four-six weeks later 14 other raters, of varying clinical experience, viewed the videos and independently scored the patients using the BBS scoring system. The original two raters also rated the patients again by viewing the videos to try to establish the intra-rater reliablity.	n=5

The studies report on a selection of measurement properties, with few presenting results on all the psychometric properties of the outcome measures being investigated, i.e. reliability, measurement error, validity and responsiveness, with internal consistency when the measure contains multiple questions or items. Whether the results represented a complete picture of the psychometric evaluation of the outcome measure will be discussed later. Measurement properties of the Locomotor Capability Index (LCI) and the modified version LCI-5, were reported by the most number of studies (n=5) (Gauthier-Gagnon et al 1998, Miller et al 2001, Rushton et al 2001, Franchignoni et al 2004 and Franchignoni et al 2007) with results available for all psychometric categories. The 2MWT was reported by three studies (Brooks et al 2001, Brooks et al 2002 and Resnik et al 2011) and again results were available for all appropriate psychometric categories (internal consistency is not required for interval scales such as the 2MWT). However, whilst the Prosthetic Evaluation Questionnaire mobility section (PEQ (ms)), Timed up and Go (TUG) and the derived L-test were also reported in 3 studies, there were no published results on the responsiveness of these outcome measures.

The Houghton Scale, the Rivermead Mobility Index (RMI), 6 minute walk test (6MWT) and the Activities-specific Balance Confidence (ABC) scale were reported by two studies each. The performance of the Houghton Scale and the RMI had been considered in all appropriate psychometric categories; with two studies presenting results for their reliability and validity, but only one for their responsiveness. The 6MWT and the ABC scale had no results presented for responsiveness. The remaining 24, of the 35, outcome measures had only one study publish results on any of their measurement properties.

4.4.3 Methodological quality results

There was 97% agreement between the two reviewers on the individual COSMIN items reviewed. There was no disagreement for the final ratings of methodological quality and therefore the third reviewer was not required.

While an overall quality rating for a study is not necessarily helpful when reviewing the utility of an outcome measure on individual measurement properties. This gives the impression that all the measurement properties have equal importance, when this may not necessarily be the case depending on what the outcome measure is to

be used for. However, an overall rating will give an indication of the methodological quality of the studies presented. The number of studies and the overall quality ratings for each study is seen in Table 4.5. Where a study presented data on one or more outcome measure or for more than one measurement property, the overall quality rating of the study was calculated using the "worst score counts". For example, Miller et al (2001) reported on reliability and measurement error for the Houghton Scale which received a fair rating, and on internal consistency which received a poor rating. Therefore, the overall methodological rating for the study was poor.

Table 4.5 Overall methodological quality rating of each study

Quality rating	Number of studies
Excellent	0
Good	9
Fair	22
Poor	5
TOTAL	36

A detailed breakdown of all the methodological quality ratings for each measurement property is given in Table 4.6. The total number of studies at the bottom of each column does not add up to the number of ratings in the measurement property category, as some studies presented data for more than one outcome measure. However, no studies were rated as having excellent methodological quality, with the majority only rated as being fair.

Table 4.6 Methodological quality ratings per measurement property

Quality Rating	Internal Consistency	Reliability	Measurement Error	Construct validity	Responsiveness
Excellent	2	0	0	0	0
Good	1	12	9	7	0
Fair	8	21	13	17	9
Poor	7	0	0	0	0
TOTAL number of studies	16	27	15	21	8

The methodological quality ratings and results for each measurement property from each study are presented by outcome measure in the following tables. Table 4.7 presents results for the measurement properties of internal consistency, reliability and measurement error, and Table 4.8 presents findings for validity, responsiveness and interpretability.

Table 4.7 Results per outcome measure: Internal consistency, reliability and measurement error

Author, year **main outcome measure (s)**	Internal consistency	Reliability	Measurement Error
Locomotor Capability Index (LCI) / LCI-5 & PPA (LCI section)			
Gauthier Gagnon et al 1998 **LCI**	Cronbachs alpha = 0.95 for original scale and 0.92 when 4 redundent items were removed. Only 1 was removed in final version. Factor analysis identified two underlying factors: relating to basic and advanced locomotor activities . **Quality: Good** % of missing items not given		
Franchignoni et al 2004 **Original LCI and revised version**	At discharge Cronbachs alpha = 0.95 Item to total correlation coefficients: 0.5 – 0.87 **Quality: Excellent**	Test-retest Reliability ICC =0.984 for both LCI & LCI-5. **Quality: Fair** Conducted during rehab phase.	Bland Altman plot revealed no systematic variations for either version **Quality: Fair** Conducted during rehab phase.
Franchignoni et al 2007 **LCI-5 and modified version LCI-10-4**	Rasch analysis highlighted redundancy of 4 items and suggested combining lowest two levels Item separation reliability = 0.98 Person– separation reliability = 0.94 for LCI-10-4 Item separation index = 7.39 Person separation index = 3.90 **Quality: Fair** Details of how missing data handled not given.		

Miller et al 2001 The Houghton Scale **PPA-(LCI)** Prosthetic Evaluation Questionnaire – mobility subscale (PEQ-ms)	Cronbach's alpha: Houghton= 0.68 PPA-LCI = 0.89 PEQ-ms = 0.95 **Quality: Poor** No factor analysis described	Test retest reliability ICC: Houghton= 0.85 PPA-LCI = 0.88 PEQ-ms = 0.77 **Quality: Fair** Details of how missing data handled not given	Standard error of measurement (SEM): Houghton= 0.93 PPA-LCI = 2.7 PEQ-ms = 1.1 **Quality: Fair** Details of how missing data handled not given
Two minute walk test (2MWT)			
Brooks et al 2002 **2MWT**		Inter and intra-rater reliability ICC > 0.98 ANOVA showed significant difference for days ($p<0.001$) but no sig for raters ($p=0.259$) **Quality: Fair** Details of how missing data handled not given.	
Resnik et al 2011 PEQ (modified version), Orthotics and Prosthetics Users' Survey (OPUS), Patient-Specific Functional Scale (PSFS), **2MWT,** 6MWT, TUG,		Test re=test reliability ICC: 2MWT = 0.83 **Quality: Good** Administrations assumed to be independent, not confirmed	Minimum Detectable Change (MDC): 2MWT = 34.3m **Quality: Good** Administrations assumed to be independent, not confirmed

Berg Balance Scale (BBS)			
Major et al 2013 **BBS**	Cronbachs alpha = 0.827 **Quality: Poor** No factor analysis performed	Inter-rater reliability ICC=0.945 **Quality: Fair** Details of how missing data handled not given.	
Wong et al 2013 **BBS**	Rasch Analysis indicated a single dimension as 70.4% was explained by the model. Person Specification Index =2.72 Item Specification Index = 4.7 **Quality: Fair** Details of how missing data handled not given		
Wong et al 2014 **BBS**		Inter-rater reliability ICC =0 .99 Intra-rater reliablity ICC = 0.99 (but only 2 raters). **Quality: Fair** Details of how missing data handled not given	
Prosthesis Evaluation Questionnaire (PEQ)			
Franchignoni et al 2007 **Original PEQ (ms) section and revised version**	Cronbachs alpha = 0.96 for both versions Item - separation reliability & Person separation reliability > 0.95 for both versions Item separation index – both versions > 7.48 Person separation index both versions > 4.15 Rasch analysis of both versions showed all items fit the model except “shower / bathe” **Quality: Fair** Details of how missing data handled not given.		

Miller et al 2001 The Houghton Scale, PPA-(LCI) **Prosthetic Evaluation Questionnaire – mobility subscale (PEQ-ms)**	Cronbach's alpha: Houghton= 0.68 PPA-LCI = 0.89 PEQ-ms = 0.95 **Quality: Poor** No factor analysis described	Test retest reliability ICC: Houghton= 0.85 PPA-LCI = 0.88 PEQ-ms = 0.77 **Quality: Fair** Details of how missing data handled not given	SEM: Houghton= 0.93 PPA-LCI = 2.7 PEQ-ms = 1.1 **Quality: Fair** Details of how missing data handled not given
Resnik et al 2011 **PEQ (modified version),** OPUS, PSFS, 2MWT, 6MWT, TUG,	PEQ overall alpha score = 0.922 Item to total scores range*: 0.69-0.79, Item to rest scores range*: 0.61-0.74 *except Shower & bathe safely **Quality: Excellent**	Test re=test reliability ICC: 2MWT = 0.83 **Quality: Good** Administrations assumed to be independent, not confirmed	MDC: 2MWT = 34.3m **Quality: Good** Administrations assumed to be independent, not confirmed
Timed Up and Go (TUG) & L-test			
Resnik et al 2011 PEQ (modified version), OPUS, PSFS, 2MWT, 6MWT**, TUG,**		Test re=test reliability ICC: TUG = 0.88 **Quality: Good** Administrations assumed to be independent, not confirmed	MDC: TUG = 3.6s **Quality: Good** Administrations assumed to be independent, not confirmed
Schoppen et al 1999 **TUG**		Spearman correlation coefficient Inter-rater reliability = 0.96 Intra-rater reliability = 0.93 **Quality: Fair** Spearman cc only presented no ICC	

Deathe et al 2005 **L-test**		Inter-rater ICC = 0.96, Intra-rater ICC = 0.97. **Quality: Fair** Details of how missing data handled not given.	Bland Altman plot suggested a ceiling effect & SEM = 3s **Quality: Fair** Details of how missing data handled not given.
Six minute walk test (6MWT)			
Lin & Bose 2008 **6MWT**		Test re-test reliability ICC: 6MWT = 0.94 **Quality: Fair** Details of how missing data handled not given.	Bland Altman plot showed no systematic variation **Quality: Fair** Details of how missing data handled not given.
Resnik et al 2011 PEQ (modified version), OPUS, PSFS, 2MWT**, 6MWT,** TUG,		Test re=test reliability ICC: 6MWT = 0.97 **Quality: Good** Administrations assumed to be independent, not confirmed	MDC:6MWT = 45m **Quality: Good** Administrations assumed to be independent, not confirmed
Activities-Specific Balance Confidence (ABC) Scale			
Miller et al 2003 **ABC Scale**	Cronbach's alpha = 0.93 Stepwise deletion of each item did not affect overall alpha **Quality: Poor** No factor analysis described	Test retest reliability ICC = 0.91 **Quality: Fair** Details of how missing data handled not given	Bland Altman plot - better repeatability for higher scores SEM = 6.3 **Quality: Fair** Details of how missing data handled not given
Sakakibari et al 2011 **ABC Scale**	Item person map showed good coverage of all levels. Cronbachs alpha range across the threshold levels = 0.75-0.94 **Quality: Fair** Details of how missing data handled not given		SEM range across the threshold levels = 0.27 – 0.4 **Quality: Fair** Details of how missing data handled not given

Houghton Scale			
Devlin et al 2004 **Houghton Scale**	Cronbachs alpha = 0.71 at discharge and 0.70 at follow up **Quality: Poor** No factor analysis applied and Cronbachs alpha applied to total score only	Test re-test reliability ICC = 0.96 Kendall r-b = 0.743 (item 1), 0.688 (item 2), 1 (item 3) Cohen's k = 0.712 (item 4a), 0.824 (item 4b), 0.453 (item 4c) **Quality: Fair** Details of how missing data handled not given	
Miller et al 2001 **The Houghton Scale** PPA-LCI, PEQ (ms)	Cronbach's alpha: Houghton= 0.68 PPA-LCI = 0.89 PEQ-ms = 0.95 **Quality: Poor** No factor analysis described	Test retest reliability ICC: Houghton= 0.85 PPA-LCI = 0.88 PEQ-ms = 0.77 **Quality: Fair** Details of how missing data handled not given	SEM: Houghton= 0.93 PPA-LCI = 2.7 PEQ-ms = 1.1 **Quality: Fair** Details of how missing data handled not given
Rivermead Mobility Index (RMI)			
Franchignoni et al 2003 **RMI**	At admission Cronbachs alpha = 0.85 Item to total correlation coefficients: 0.33-0.74 **Quality: Fair** Details of how missing data handled not given		
Ryall et al 2003 **RMI** Timed 10m walk (from standing)	Rasch analysis: Person separation =1.66 Item Separation 7.99 **Quality: Fair** Details of how missing data handled not given	Intra-rater reliability: ICC= 0.99 Inter-rater: Kappa coefficient ranged 0.78 - 1 **Quality: Fair** Details of how missing data handled not given.	

Assessment of Daily Activity Performance in Transfemoral amputees (ADAPT)			
Theeven et al 2010 **ADAPT**		Test re-test reliability Pearson's correlation r>0.8 except for 1 station (strip bedclothes) **Quality: Fair** Details of how missing data handled not given	Bland Altman plot shows high levels of agreement with no specific trends or biases **Quality: Fair** Details of how missing data handled not given
Climbing Stair Questionnaire			
de Laat et al 2010 **Climbing stair Questionnaire**		Test re-test reliability ICC 0.79; 95% CI, (0.57–0.90). **Quality: Good** Postal - test conditions assumed identical.	Bland Altman plot: Overall acceptable agreement, but large differences for two subjects in mid-range **Quality: Good** Postal - test conditions assumed identical.
DSF-84			
Monteiro et al 2013 **DSF-84**	Cronbachs alpha range across 5 domains = 0.52-0.89 **Quality: Poor** No factor analysis applied	Inter-rater ICC > 0.99 all domains **Quality: Fair** Details of how missing data handled not given	
Frenchay Activities Index (FAI)			
Miller et al 2004 **FAI**	Cronbachs alpha = 0.81 Factor analysis confirmed 3 factors **Quality: Fair** Details of how missing data handled not given	Test-retest reliability ICC=0.79 **Quality: Fair** Details of how missing data handled not given	Bland Altman plot large variations > 2SDs

Functional Measure for Amputees (FMA)			
Callaghan et al 2002 **FMA**		Test-test reliability ICC: continuous - 0.64 – 0.96 categorical data 20 - 70% **Quality: Good** Postal - test conditions assumed identical and patients stable.	
Goal Attainment Scaling (GAS) & Barthel Index			
Rushton et al 2002 **GAS**		Inter-rater reliability ICC = 0.67 **Quality: Fair** Details of how missing data handled not given & early rehab so subjects may have changed	
Harold Wood-Stanmore Mobility Scale, Handicap Mobility Scale, Handicap Physical Independence Scale.			
Gardiner et al 2002 **Harold Wood -Stanmore Mobility Scale, Handicap Mobility Scale, Handicap Physical Independence Scale.**		Inter-rater reliability ICC: HW-S adm = 1, d/c = 0.83 Handi (mob) adm = 0.49, d/c = 0.83 Handi (phys) adm = 0.15, d/c = 0.69 **Quality: Fair** Details of how missing data handled not given.	
Hill Assessment Index (HAI)			
Highsmith et al 2013 **HAI**		Inter-rater reliability ICC: test 1= 0.97 test 2 = 0.99 **Quality: Good** Spearman's Rho used to calculate ICC	

One leg stand test (OLST)			
Kristensen et al 2014 **OLST**		Inter-rater reliability ICC=0.87 **Quality: Fair** Details of how missing data handled not given.	SEM and single patient Smallest Real Difference (SRD): *OLST 10s*: SEM=2.92 SRD=8.09 *OLST 20s*: SEM= 5.84 SRD= 16.18 *OLST 40s*: SEM 11.68 SRD= 32-36 **Quality: Fair** Details of how missing data handled not given.
Orthotics and Prosthetics Users' Survey (OPUS), Patient-Specific Functional Scale (PSFS)			
Resnik et al 2011 PEQ (modified version), **OPUS, PSFS,** 2MWT, 6MWT, TUG,		Test re=test reliability ICC: OPUS = 0.67 PSFS = 0.83 **Quality: Good** Administrations assumed to be independent, not confirmed	MDC: OPUS = 10.3pt PSFS = 11.2pt **Quality: Good** Administrations assumed to be independent, not confirmed
Physiological Cost Index (PCI)			
Hagberg et al 2010 **PCI**		Intra-rater reliability ICC = 0.966 **Quality: Good** no model of ICC described	Bland Altman plot showed acceptable agreement SDC = 0.116, coeffiecient of variation = 20% **Quality: Good** no model of ICC described

Prosthetic Observational Gait Score (POGS)			
Hillman et al 2010 **POGS**		Intra-rater reliability coefficient of repeatability (CoR) range 1.5-4.6 Inter-rater CoR range 6-5.9 % agreement of each item of the score range 50-87. Kappa statistics range -0.03-0.6 **Quality: Fair** Details of how missing data handled not given	.
Prosthetic Profile of the Amputee (PPA)			
Gauthier-Gagnon et al 1994 **PPA**		Test-retest reliability Continuous data: ICC 0.8 – 0.92 Categorical data: k=0.46 – 0.86 **Quality: Good** Postal - test conditions assumed identical	
Rising and sitting down Questionnaire			
de Laat et al 2011 **Rising and sitting down Questionnaire**		Test re-test reliability ICC 0.83, 95% CI (0.65 - 0.93) **Quality: Good** Postal - test conditions assumed identical.	Bland Altman plot, acceptable overall but large differences for two subjects in mid-range SDC = 18.6% **Quality: Good** Postal - test conditions assumed identical

Special Interest Group in Amputee Medicine (SIGAM) mobility grade			
Ryall et al 2003 **SIGAM**	Rasch analysis: Person separation = 2.6 Item Separation 7.99 **Quality: Fair** Details of how missing data handled not given	Test re-test ICC = 0.79 Kappa coefficient = 0.86 **Quality: Fair** Details of how missing data handled not given	
Timed 10m walk test (from standing)			
Ryall et al 2003 RMI **Timed 10m walk (from standing)**		Intra-rater reliability ICC= 0.99 Inter-rater Kappa coefficient ranged 0.78 - 1 **Quality: Fair:** Details of how missing data handled not given	Bland Altman plot indicated high level of agreement at the two time pts **Quality: Fair:** Details of how missing data
Walking Questionnaire			
de Laat et al 2012 **Walking Questionnaire**		Test re-test reliability ICC 0.73, 95% CI (0.43 - 0.88) **Quality: Good** Postal - test conditions assumed identical.	Bland Altman plot, acceptable overall but large differences for one subject in mid-range **Quality: Good** Postal - test conditions assumed identical.

Sixteen studies looked at the **internal consistency** of 9 outcome measures. The majority of the studies reported Cronbachs alpha (10/16), while factor analysis was performed in 11 out of the 16 studies to investigate the dimensionality of the measure. This demonstrates that both the Classical Test Theory approach and the Item Test Theory had been used by the authors.

Twelve studies looked at **inter-rater reliability** in thirteen different outcome measures. Two studies used video footage of the same patient, with the remaining studies testing the same patient on two different occasions with time intervals varying between 2 hours and 6 weeks.

Twenty studies looked at the **intra-rater reliability** reporting results on 21 outcome measures. Most studies utilised the test-retest study-design with the same patient and time intervals ranged between the same day and up to 2 weeks. This extended to 6 weeks when video footage of the patient was employed (Hillman et al 2010).

Where intra-class correlations (ICCs) had been presented, the particular model was mentioned in 13 out of the 27 studies. Where the model was reported the reason for choosing a particular model was often not stated. Within the COSMIN checklist, this lack of clarity over the ICC model used, results in a lower quality rating.

When testing reliability, it is essential to consider any day to day variation in the health (clinical) status of a participant that may impact on test conditions and increase the random measurement variation, and this was detailed in five studies. Stability of the participant's clinical condition was assumed in the remainder where no details had been given. A "good" rating was given in these cases, unless the phase of rehabilitation and time interval caused concern that a change in the patient's clinical condition may have occurred.

The same scrutiny was applied when judging the item "test conditions". Any variation in how the measure was implemented e.g. in clinic and a subsequent postal delivery for the first and second administration of a self-reported measure, was considered "not similar" and the quality rating was reduced. In many cases, the "test conditions" were only assumed to be identical, i.e. no details had been given, and therefore only a "good" rating could be given.

Means and standard deviations (SD) were offered in most studies reporting on reliability. However, measurement error or agreement parameters, e.g. standard error of measurement (SEM), limit of agreement (LoA), minimal detectable change (MDC) or equivalent, were not presented in 13 out of the 27 studies that reported reliability results.

Of the 14 studies that had reported **measurement error** parameters, five presented SEM results for six different outcome measures: L-test; one leg stand test; Houghton Scale; PEQ (ms); PPA (LCI); and ABC scale. In addition, six studies described LoA for: PCI; Frenchay Activities Index; ABC scale; 10m timed walk (from standing); and ADAPT. Values for MDC, or equivalent using synonymous terms, were presented for outcome measures in four studies (see chapter 2 for a full definition of the variety of terms used to describe similar properties). Minimal detectable change values were calculated for the PEQ (ms), OPUS, PSFS, 2 & 6 MWT and TUG by Resnik (2011); the smallest detectable difference (SDD) was calculated for the Rising and Standing Questionnaire by de Laat et al (2011); the smallest detectable change (SDC) was calculated for the PCI by Hagberg et al (2011); and the smallest real difference (SRD) was calculated for the one leg stand test (Kristensen et al 2014). All the authors used the SEM to calculate these values except Hagberg et al (2011) who had used the intra-individual standard deviation (IISD). All parameters had been reported using 95% CI, with the exception of those in the study by Resnik et al (2011), where 90% CI had been reported. There was one study that presented SEM values without presenting any reliability data (Sakakibara et al 2011).

Table 4.8 Methodological quality ratings and results (level of evidence) for validity / responsiveness / interpretability per outcome measure.

Author, year **Main outcome measure (s)**	Construct Validity	Responsiveness	Interpretability
Locomotor Capability Index (LCI) / LCI-5 & PPA (LCI section)			
Franchignoni et al 2004 **Original LCI and revised version**	Spearmans *p* values 2MWT = 0.667 (LCI), 0.708 (LCI-5), RMI = 0.752 (LCI), 0.757(LCI-5), FIM = 0.617(LCI), 0.622 (LCI-5) **Quality: Good** Only two aprori hypotheses presented.	Wilcoxon signed rank showed significant differences for both LCI and LCI-5 during rehabilitation period (mean 36 days) Effect size = 1.09 (LCI) & 1.40 (LCI-5) **Quality: Fair** Effect size	The percentage of participant who had the highest possible scores reflected a change in the ceiling effect reported for the original version compared to the LCI-5
Franchignoni et al 2007 **LCI-5 and modified version LCI-10-4**	Spearmans r_s = 0.77 showed strong correlation with the PEQ(MS) **Quality: Fair** Details of how missing data handled not given.		Study population relatively young: mean age = 54 High prevalence of traumatic amputees (56%)
Miller et al 2001 The Houghton Scale **PPA-(LCI)** Prosthetic Evaluation Questionnaire – mobility subscale (PEQ-ms)	All hypothesized relationships confirmed by Pearson's r: Houghton/PPA-LCI= 0.59 Houghton/PEQ-ms= 0.55 PPA-LCI/PEQ-ms= 0.83 **Quality: Fair** Details of how missing data handled not given		

Rushton et al 2002 Goal Attainment Scaling (GAS) **Comparator results LCI**		Effect size LCI = 3.7 Relative efficiencies Houghton vs LCI = 4.7 **Quality: Fair** Details of how missing data handled not given. Interim period not described. Effect size	
Two minute walk test (2MWT)			
Brooks et al 2001 **2MWT**	Pearsons correlation r=0.22 between distance walked and SF-36 prior to d/c r=0.48 at follow up r=0.49 between Houghton total score and distance walked at d/c **Quality: Fair** Details of how missing data handled not given. Houghton and SF-36 used in the absence of a gold standard measure of physical function for pwLLAs. SF-36 not validated in pwLLAs. Houghton scale measures prosthetic use, not mobility.	The distances walked indicated a change during rehabilitation. ANOVA , p<0.001 **Quality: Fair** Details of how missing data handled not given. The interim period was not described. ANOVA calculated on mean differences.	Retrospective study. Distances walked did indicate a change but no SDC or MIC was calculated.
Berg Balance Scale (BBS)			
Major et al 2013 **BBS**	Spearman's rank correlation with: ABC scale=0.634, PEQ(ms) = 0.584, FAI = 0.607, 2MWT = 0.675, L-test = -0.802 **Quality: Fair** Details of how missing data handled not given		

Prosthesis Evaluation Questionnaire (PEQ)			
Franchignoni et al 2007 **PEQ (ms) and revised version**	All hypothesized relationships confirmed by Spearmans *r* LCI= 0.78 **Quality: Fair** Details of how missing data handled not given.		Study population relatively young: mean age = 54 High prevalence of traumatic amputees (56%)
Miller et al 2001 The Houghton Scale PPA-(LCI) **PEQ-(ms)**	All hypothesized relationships confirmed by Pearson's *r*: Houghton/PPA-LCI = 0.59 Houghton/PEQ-(ms) = 0.55 PPA-LCI/PEQ-(ms) = 0.83 **Quality: Fair** Details of how missing data handled not given		
Timed Up and Go (TUG) & L-test			
Schoppen et al 1999 **TUG**	All hypothesized relationships confirmed by Spearmans *r*: Sickness Impact Profile (SIP68) mobility control =0.46 mobility range = 0.36 Groningen Activity Restriction Scale (GARS) = 0.39 **Quality: Good** Spearman only presented, no means or SD		Elderly vascular subjects only therefore results not generalisable.

Deathe et al 2005 **L-test**	Pearsons correlations (r = 0.22 – 0.93) with other mobility measures recorded. **Quality: Fair** Details of how missing data handled not given.		Bland Altman plot suggested a ceiling effect in the L-test however 3 recorded ceiling effect for L-Test not TUG and 14 recorded ceiling effect for TUG not L-Test. McNemar Test showed 66 did not demonstrate ceiling effects.
Six minute walk test (6MWT)			
Lin & Bose 2008 **6MWT**	Pearson's correlation: TUG r=0.76, One leg balance test (eyes open) r=0.00, (eyes closed) r=0.42 **Quality: Fair** Details of how missing data handled not given.		
Activities-Specific Balance Confidence (ABC) Scale ABC scale			
Miller et al 2003 **ABC scale**	All hypothesized relationships confirmed by Pearson's r: 2MWT= 0.72 TUG = -0.70 **Quality: Fair** Details of how missing data handled not given		ABC scale did not discriminate between groups according to amputation level.
Houghton Scale			
Devlin et al 2004 **Houghton Scale**	All hypothesized relationships confirmed by Pearsons correlations (r = 0.24 - 0.65) **Quality: Fair** Details of how missing data handled not given. And aprori hypotheses not presented.	Wilcoxon signed rank showed significant differences for mean score changes from discharge to follow-up Effect size = 0.6 3mth period after d/c **Quality: Fair** Effect size	

Miller et al 2001 **The Houghton Scale** PPA-(LCI) PEQ-(ms)	All hypothesized relationships confirmed by Pearson's r: Houghton/PPA-LCI = 0.59 Houghton/PEQ-(ms) = 0.55 PPA-LCI/PEQ-(ms) = 0.83 **Quality: Fair** Details of how missing data handled not given		
Rivermead Mobility Index (RMI)			
Franchignoni et al 2003 **RMI**	Spearman rank correlations (r_s = 0.69 – 0.7) with other measures recorded **Quality: Fair** Details of how missing data handled not given	Effect size = 1.35 calculated for prosthetic training period (time not defined) **Quality: Fair** Effect size	RMI appears more useful for epidemiological studies than at an individual level on the basis of this study.
Ryall et al 2003 **RMI** Timed 10m walk (from standing)	Spearman's Rho TWT = -0.58 **Quality: Fair** Details of how missing data handled not given. Apriori hypotheses not fully described.		
Assessment of Daily Activity Performance in Transfemoral amputees (ADAPT)			
Theeven et al 2010 **ADAPT**	Mann Whitney demonstrated significant differences in performance times for 6/9 (75%) of the test stations **Quality: Fair** Details of how missing data handled not given		ADAPT appears to be able to differentiate between groups of users by their prosthetic componentry.

Climbing Stair Questionnaire			
de Laat et al 2010 **Climbing stair Questionnaire**	2/10 hypotheses rejected (> 75% confirmed) - validity considered good. **Quality: Good** Magnitude of expected differences not stated a priori		Recommended for group comparisons but not individual
Frenchay Activities Index (FAI)			
Miller et al 2004 **FAI**	All hypothesized relationships confirmed by Pearson's r 2MWT =0.526 TUG = -0.486 PEQ(ms) =0.385 ABC =0.505 **Quality: Fair** Details of how missing data handled not given		Reliability is adequate to detect group but not individual level differences
Goal Attainment Scaling (GAS)			
Rushton et al 2002 **GAS** Barthel Index (comparator only)	All hypothesized relationships confirmed by Pearson's r : Barthel Index = 0.44 LCI = 0.35 **Quality: Fair** Details of how missing data handled not given	Effect size = 6.5 Relative efficiencies vs Barthel Index=3.1 Vs LCI = 4.7 **Quality: Fair** Details of how missing data handled not given & interim period not described. Effect size	

ICF Checklist			
Kohler et al 2011 **ICF Checklist**		Wilcoxon signed rank tests indicated a functional significant deterioration immediately after amputation with a gradual improvement over the next 3 weeks. **Quality: Fair** No apriori hypotheses formulated	
Office of Population Censuses and Surveys (OPCS) Scale, Functional Independence Measure (FIM), Amputee Activity Score (AAS)			
Panesar et al 2001 **OPCS Scale,** **FIM,** **AAS**	Kendall's coefficient of concordance X^2 shown with each other at each time point, and with LCI confirming hypotheses: adm = 65.79 d/c = 71.8 follow-up = 72.81 LCI = 57.28 **Quality: Good** Expected direction and magnitude of hypotheses not stated	All scores showed statistically significant changes ($p<0.00001$) between adm & d/c Kendalls correlation coefficient z: OPCS = 6.25 FIM = 5.37 AAS = 5.39 **Quality: Fair** No description of interim period.	
Prosthetic Profile of the Amputee (PPA)			
Gauthier-Gagnon et al 1994 **PPA**	Two hypotheses confirmed using Pearson's product-moment correlation coefficients for all three constructs of PPA and RNL index: (0.30 – 0.73) **Quality: Good** Postal - test conditions assumed identical		

Rising and sitting down Questionnaire			
de Laat et al 2011 **Rising and sitting down Questionnaire**	1/10 hypotheses rejected (> 75% confirmed) **Quality: Good** Magnitude of expected differences not stated a priori		Recommended for group comparisons and not individual.
Special Interest Group in Amputee Medicine (SIGAM) mobility grade			
Ryall et al 2003 **SIGAM**	Significant differences between timed walk tests and non-adjacent grades were found. **Quality: Fair** Details of how missing data handled not given. Apriori hypotheses not fully described.	Effect size = 10.66 **Quality: Fair** Details of how missing data handled not given. Apriori hypotheses not fully described. Effect size	
Walking Questionnaire			
de Laat et al 2012 **Walking Questionnaire**	1/11 hypotheses rejected (> 75% confirmed) **Quality: Good** Magnitude of expected differences not stated a priori		Recommended for group comparisons and not individual.

Gold standard measures are not well established within the field of rehabilitation and as a consequence criterion validity was not reported. Hypothesis testing was used in all of the studies (n=21) aiming to report on **construct validity,** with none reporting criterion validity. Where correlations with comparator outcome measures were presented, they were mostly either Pearson's product-moment correlation (10 studies) or Spearman's rank correlations (9 studies). In one study, Kendall's k correlations were presented and one study used Rasch analysis to provide additional information on the structural validity of the outcome measure.
The quality of the hypotheses that had been presented, were variable. Most stated *a priori* hypotheses with expected relationships to comparator outcome measures that purported to be measuring the same construct(s). Where explicit hypotheses had not been stated, the intention was often easy to interpret, but this still reduced the methodological quality level for that property to "fair".

Responsiveness, or the ability of the outcome measure to measure a change in the clinical condition, was reported in eight studies on 11 outcome measures: 2MWT; Houghton scale; Rivermead Mobility Index; LCI; ICF checklist; FIM; OPCS; Amputee Activity Score; Goal Attainment Scaling; Barthel Index; and SIGAM. Analysis of variance (ANOVA), effect size, Wilcoxon signed rank test, Kendall's correlation coefficient and Relative Efficiency results had been presented to demonstrate the responsiveness of the OM under study and to detect change in the clinical condition (in a variety of settings) using a longitudinal study design. While these eight studies reported either statistical significance of the differences or the magnitude of change, or both, no authors reported the "clinical" responsiveness of the outcome measures under study, i.e. whether any detected changes in scores reflected a true clinical change in condition or function, using either a subjective or gold standard anchor. Minimally important clinical differences (MICD) were not reported by any of the studies included in this review.

4.4.4 Levels of evidence of the results

The level of evidence ratings was awarded by examining the results obtained in each study for each measurement-property against the standards described in Terwee et al (2007); see Table 4.1 for full details of the standards. Results for 106 measurement properties were reported for 35 outcome measures across the 36 studies reviewed, and all except three were deemed to be positive ''+". The

Orthotics and Prosthetics Users' Survey (OPUS) presented in the study by Resnik et al (2011), the Goal Attainment Scaling (GAS) OM (Rushton et al 2002) and the continuous items in the PPA (Gauthier-Gagnon et al 1994), all presented results that were considered to have a negative rating for reliability, i.e. ICC or weighted Kappa <0.70.

There were twenty studies where the evidence had been awarded a positive rating but the final rating was changed to indeterminate "?" due to small sample sizes, i.e. < 50 participants. The remaining indeterminate ratings were due to: no factor analysis and/or no Cronbach's alpha reported for internal consistency, both of which are required for a positive result; no comment made on the limits of agreement when reporting on measurement error which is required; or a poor choice of statistical analysis when reporting responsiveness, i.e. no ROC analysis or reference to any minimal importance change (MIC) values. All the evidence ratings together with a the methodological quality ratings are presented in Table 4.9.

Table 4.9 Summary of quality ratings and levels of evidence

Author, year **Main outcome measure(s)**	Internal consistency	Reliability	Measurement Error	Construct Validity	Responsiveness
Locomotor Capability Index (LCI) / LCI-5 & PPA (LCI section)					
Gauthier Gagnon et al 1998 **LCI**	**Quality: Good** (% of missing items not given) **Evidence: +**				
Franchignoni et al 2004 **Original LCI and revised version**	**Quality: Excel** **Evidence: +**	**Quality: Fair** (Conducted during rehab phase) **Evidence: ?** (sample size)	**Quality: Fair** (Conducted during rehab phase) **Evidence: ?** (sample size)	**Quality: Good** (Only two aprori hypotheses presented) **Evidence: +**	**Quality: Fair** (Effect size) **Evidence: ?** (choice of analysis)
Franchignoni et al 2007 **LCI-5 and modified version LCI-10-4**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**			**Quality: Fair** Details of how missing data handled not given. **Evidence: +**	
Miller et al 2001 **PPA-(LCI)**	**Quality: Poor** (No factor analysis described) **Evidence: ?** (no factor analysis)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (no comment on LoA)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	

Rushton et al 2002 Comparator results **LCI**					**Quality: Fair** (Details of how missing data handled not given & interim period not described & effect size) **Evidence: ?** (choice of analysis)
Two minute walk test (2MWT)					
Brooks et al 2001 **2MWT**				**Quality: Fair** (Details of how missing data handled not given. Houghton and SF-36 were used in the absence of a gold standard measure of physical function for pwLLAs. SF-36 not validated in pwLLAs. Houghton Scale measures prosthetic use, not mobility) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given. The interim period was not described. ANOVA calculated on mean differences) **Evidence: ?** (choice of analysis)
Brooks et al 2002 **2MWT**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)			

Resnik et al 2011 **2MWT**		**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (sample size)	**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (no comment on LoA)		
Berg Balance Scale (BBS)					
Major et al 2013 **BBS**	**Quality: Poor** (No factor analysis performed) **Evidence:?** (no factor analysis)	**Quality: Fair**. (Details of how missing data handled not given) **Evidence: +**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	
Wong et al 2013 **BBS**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (no Cronbachs alpha)				
Wong et al 2014 **BBS**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)			

PEQ original and mobility scale					
Franchignoni et al 2007 **PEQ (ms) and revised version**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**			**Quality: Fair** (Details of how missing data handled not given) **Evidence+**	
Miller et al 2001 **PEQ-(ms)**	**Quality: Poor** (No factor analysis described) **Evidence: ?** (no factor analysis)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (no comment on LoA)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	
Resnik et al 2011 **PEQ (modified version),**	**Quality: Excellent** **Evidence: ?** (sample size)	**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (sample size)	**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (no comment on LoA)		
Timed Up and Go (TUG) & L-test					
Resnik et al 2011 **TUG**		**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (sample size)	**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (no comment on LoA)		

Schoppen et al 1999 **TUG**		**Quality: Fair** (Spearman cc only presented no ICC) **Evidence: ?** (sample size)		**Quality: Good** (Spearman cc only presented no ICC) **Evidence: ?** (sample size)	
Deathe et al 2005 **L-test**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: inter +, intra ?** (sample size)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	
Six minute walk test (6MWT)					
Lin & Bose 2008 **6MWT**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	
Resnik et al 2011 **6MWT**		**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (sample size)	**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (no comment on LoA)		

Activities-Specific Balance Confidence (ABC) Scale					
Miller et al 2003 **ABC scale**	**Quality: Poor** (No factor analysis described) **Evidence: ?** (no factor analysis)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	
Sakakibari et al 2011 **ABC Scale**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (no comment on LoA)		
Houghton Scale					
Devlin et al 2004 **Houghton Scale**	**Quality: Poor** (No factor analysis applied and Cronbachs alpha applied to total score only) **Evidence: ?** (no factor analysis)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)		**Quality: Fair** (Details of how missing data handled not given. And aprori hypotheses not presented.) **Evidence: ?** (no hypotheses)	**Quality: Fair** (Effect size) **Evidence: ?** (choice of analysis)
Miller et al 2001 **The Houghton Scale**	**Quality: Poor** (No factor analysis described) **Evidence:?** **(**no factor analysis)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (no comment on LoA)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	

Rivermead Mobility Index (RMI)					
Franchignoni et al 2003 **RMI**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**			**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Effect size) **Evidence: ?** (choice of analysis)
Ryall et al 2003 **RMI**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (no Cronbachs alpha)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given. Apriori hypotheses not fully described.) **Evidence: +**	
Assessment of Daily Activity Performance in Transfemoral amputees (ADAPT)					
Theeven et al 2010 **ADAPT**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	
Climbing Stair Questionnaire					
de Laat et al 2010 **Climbing stair Questionnaire**		**Quality: Good** (Postal - test conditions assumed identical) **Evidence: ?** (sample size)	**Quality: Good** (Postal - test conditions assumed identical) **Evidence: ?** (sample size)	**Quality: Good** (Magnitude of expected differences not stated a priori) **Evidence: +**	

DSF-84					
Monteiro et al 2013 **DSF-84**	**Quality: Poor** (No factor analysis described) **Evidence: ?** (no factor analysis)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)			
Frenchay Activities Index (FAI)					
Miller et al 2004 **FAI**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	
Functional Measure for Amputees (FMA)					
Callaghan et al 2002 **FMA**		**Quality: Good** (Postal - test conditions assumed identical and patients stable) **Evidence: +**			
Goal Attainment Scaling (GAS)					
Rushton et al 2002 **GAS**		**Quality: Fair** (Details of how missing data handled not given & early rehab so subjects may have changed) **Evidence: -**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	**Quality: Fair** (Details of how missing data handled not given & interim period not described. Effect size) **Evidence: ?** (choice of analysis)

Harold Wood -Stanmore Mobility Scale, Handicap Mobility Scale, Handicap Physical Independence Scale.					
Gardiner et al 2002 **Harold Wood -Stanmore Mobility Scale, Handicap Mobility Scale, Handicap Physical Independence Scale.**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)			
Hill Assessment Index (HAI)					
Highsmith et al 2013 **HAI**		**Quality: Good** (Spearman's Rho used to calculate ICC) **Evidence: ?** (sample size)			
ICF Checklist					
Kohler et al 2011 **ICF Checklist**					**Quality: Fair** (No apriori hypotheses formulated) **Evidence: ?** (choice of analysis)
Office of Population Censuses and Surveys (OPCS) Scale, Functional Independence Measure (FIM), Amputee Activity Score (AAS)					
Panesar et al 2001 **OPCS Scale,** **FIM,** **AAS**				**Quality: Good** (Expected direction and magnitude of hypotheses not stated) **Evidence: ?** (sample size)	**Quality: Fair** (No description of interim period.) **Evidence: ?** (choice of analysis)

One leg stand test (OLST)					
Kristensen et al 2014 **OLST**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (no comment on LoA)		
Orthotics and Prosthetics Users' Survey (OPUS), Patient-Specific Functional Scale (PSFS)					
Resnik et al 2011 **OPUS,** **PSFS**		**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: -**	**Quality: Good** (Administrations assumed to be independent, not confirmed) **Evidence: ?** (no comment on LoA)		
Physiological Cost Index (PCI)					
Hagberg et al 2010 **PCI**		**Quality: Good** (no model of ICC described) **Evidence: ?** (sample size)	**Quality: Good** (no model of ICC described) **Evidence: ?** (sample size)		
Prosthetic Observational Gait Score (POGS)					
Hillman et al 2010 **POGS**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)	.		

Prosthetic Profile of the Amputee (PPA)					
Gauthier-Gagnon et al 1994 **PPA**		**Quality: Good** (Postal - test conditions assumed identical) **Evidence: Cont +, Cat –**		**Quality: Good** (Postal - test conditions assumed identical) **Evidence: +**	
Rising and sitting down Questionnaire					
de Laat et al 2011 **Rising and sitting down Questionnaire**		**Quality: Good** (Postal - test conditions assumed identical) **Evidence: ?** (sample size)	**Quality: Good** (Postal - test conditions assumed identical) **Evidence: ?** (sample size)	**Quality: Good** (Magnitude of expected differences not stated a priori) **Evidence: +**	
Special Interest Group in Amputee Medicine (SIGAM) mobility grade					
Ryall et al 2003 **SIGAM**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (no Cronbachs alpha)	**Quality: Fair** (Details of how missing data handled not given) **Evidence: ?** (sample size)		**Quality: Fair** (Details of how missing data handled not given. Apriori hypotheses not fully described.) **Evidence: ?** (no hypotheses)	**Quality: Fair** (Details of how missing data handled not given. Apriori hypotheses not fully described. Effect size) **Evidence: ?** (choice of analysis)
Timed 10m walk test (from standing)					
Ryall et al 2003 **Timed 10m walk (from standing)**		**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**	**Quality: Fair** (Details of how missing data handled not given) **Evidence: +**		

Walking Questionnaire					
de Laat et al 2012 **Walking Questionnaire**		**Quality: Good** (Postal - test conditions assumed identical) **Evidence: ?** (sample size)	**Quality: Good** (Postal - test conditions assumed identical) **Evidence: ?** (sample size)	**Quality: Good** (Magnitude of expected differences not stated a priori) **Evidence: +**	

4.4.5 Best evidence synthesis

Assessment of the strength of evidence was determined using the levels of evidence together with the methodological quality against criteria noted in table 4.10**.**

Table 4.10 Strength of evidence for the measurement properties (based on the Cochrane Back Review Group 2003 (van Tulder et al. 2003) adapted from Terwee et al 2007)

Level	Rating	Criteria
Strong	+++ or ---	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	++ or --	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality OR in one study of good methodological quality
Limited	+ or -	One study of fair methodological quality
Conflicting	±	Conflicting findings
Unknown	?	Only studies of poor methodological quality
+ = positive results; ? =indeterminate results; - =negative results		

The full results of the best evidence synthesis are detailed in Table 4.11, with a summary of the strength of evidence presented for physical function outcome measures in pwLLAs presented in Table 4.12.

Table 4.11 Synthesis of best evidence

Outcome Measure (number of studies)	Measurement Parameter Strength of Evidence				
	Internal Consistency	**Reliability**	**Measurement Error**	**Construct Validity**	**Responsiveness**
LCI / LCI-5 / PPA-LCI section (5)	**Strong positive**	**Limited positive**	**Unknown**	**Strong positive**	**Unknown**
2MWT (3)	**N/A**	**Unknown**	**Unknown**	**Limited positive**	**Unknown**
BBS (3)	**Unknown**	**Limited positive**		**Unknown**	
PEQ /PEQ (ms) (3)	**Limited positive**	**Limited positive**	**Unknown**	**Moderate positive**	
TUG (3)	**N/A**	**Unknown**	**Unknown**	**Unknown**	
L-Test (1)	**N/A**	**Limited + & -**	**Limited positive**	**Limited positive**	
6MWT (2)	**N/A**	**Unknown**	**Unknown**	**Unknown**	
ABC Scale (2)	**Limited positive**	**Limited positive**	**Limited positive**	**Limited positive**	
Houghton Scale (2)	**Unknown**	**Limited positive**	**Unknown**	**Limited positive**	**Unknown**
RMI (2)	**Limited positive**	**Limited positive**	**Limited positive**	**Moderate positive**	**Unknown**
ADAPT (1)		**Unknown**	**Unknown**		
Climbing Stair Questionnaire (1)		**Unknown**	**Unknown**	**Moderate positive**	
DSF-84 (1)	**Unknown**	**Unknown**			
FAI (1)	**Limited positive**	**Limited positive**	**Limited positive**	**Limited positive**	
FMA (1)		**Moderate positive**			
GAS (1)		**Unknown**		**Unknown**	**Unknown**

Barthel Index (comparator only) (1)					**Unknown**
Harold Wood/Stanmore Mobility Scale (1)		**Unknown**			
Handicap Mobility Scale (1)		**Unknown**			
Handicap Physical Scale (1)		**Unknown**			
Hill Assessment Index (1)		**Unknown**			
ICF Checklist (1)					**Unknown**
OPCCS, (1)				**Unknown**	**Unknown**
FIM (1)				**Unknown**	**Unknown**
AAS (1)				**Unknown**	**Unknown**
One leg stand test (1)		**Unknown**	**Unknown**		
OPUS (1)		**Unknown**	**Unknown**		
PSFS (1)		**Unknown**	**Unknown**		
PCI (1)		**Unknown**	**Unknown**		
POGS (1)		**Unknown**			
PPA (1)		**Moderate + & -**		**Moderate positive**	
Rising and sitting down Questionnaire (1)		**Unknown**	**Unknown**	**Moderate positive**	
SIGAM (1)	**Unknown**	**Unknown**		**Unknown**	**Unknown**
Timed 10m walk test (1)	**N/A**		**Limited positive**	**Moderate positive**	
Walking Questionnaire (1)		**Unknown**	**Unknown**		

Table 4.12 Summary of evidence strength

Strength of evidence	Measurement Properties				
	Internal Consistency	Reliability	Measurement Error	Construct Validity	Responsiveness
Strong +	1				
Moderate +		2		7	
Moderate -		1			
Limited +	4	8	5	5	
Limited -		1			
Unknown	4	19	14	8	11
Total	***9***	***31****	***18***	***20***	***11***

** strength of evidence results for PPA were different for categorical and ordinal data and for Ltest they were different for intra- and inter-rater reliability, therefore the number of outcome measures does not match the number studied (see table 4.2)*

With the exception of construct validity and internal consistency, the majority of evidence for reliability, measurement error and responsiveness is unknown. Eight outcome measures have strong or moderate positive evidence in one or more measurement properties: LCI/LCI-5, PEQ, RMI, Climbing Stair Questionnaire, FMA, PPA, Rising and Sitting Down Questionnaire and the Walking Questionnaire. Five outcome measures have limited evidence to support positive results for measurement error: L-Test, ABC scale, RMI, FAI and 10m walk test. Evidence for responsiveness for any outcome measure in this population was either absent or unknown.

4.5 Discussion

The aims of this systematic review were to: a) examine the published literature investigating outcome measures used to measure physical function of lower limb amputees walking with a prosthesis; b) provide a methodological quality-rating for each psychometric measurement property presented; and c) report the level of evidence obtained for each outcome measure.

Synthesis of the results demonstrated a scarcity of high quality evidence. The implications of this will be discussed below, taking into account the context of the population pwLLA and also the intended use of an outcome measure.

4.5.1 Quantity and quality of evidence

Evidence is either not present or unknown for at least one measurement property, for each of the many outcome measures reported in this review. The lack of data presented to support the responsiveness of these outcome measures is especially noted. Results from previous reviews involving pwLLAs have commented on the limited evidence of responsiveness (Sawers and Hafner 2013, Condie et al. 2006, Deathe et al. 2009, Hebert et al. 2009).

The review carried out by Hawkins et al in 2014 did not make any comment on the evidence on responsiveness, but they did note that several OMs were rated +++, as per the Johnstone and Graves criteria (Johnston and Graves 2008). The definition of +++ is:

> "Adequately/reasonably valid for the main defined purpose. (Widely used outside of *population of interest*, with formal studies/use in *population of interest*.). OK to use in studies, although checking of assumptions or small improvements may be desirable to further improve the measure (eg, classical measures would benefit from item-response theory or Rasch analysis) "

This definition does leave some element of doubt over the completeness of the measures and while the authors comment in their discussion that "there are a number of validated instruments that are of high quality and are widely used", the results of this current systematic review do not concur with the label of high quality. Within the Johnstone and Graves criteria, used in this review as well as the Deathe et al (2009) and Hebert et al (2009) reviews, there is no requirement to qualify measures of "sensitivity to change" (assessed within validity), merely to report significant differences or medium to large effect size. Responsiveness of the measures may therefore be under reported in these reviews. Also "overall" quality is presented, not per measurement property as in COSMIN.

Heineman et al (2014) presented a narrative review of the outcome instruments for prosthetics. In their abstract they stated “It is encouraging that responsiveness of measures is often reported, as this information is needed to improve clinical utility.” However with no quality assessment of the studies or grading of the level or strength of evidence presented, there appears to be no basis for this statement.

Systematic reviews in other patient populations have also found a similar lack of evidence for responsiveness and measurement error in walking tests for stroke survivors (Scrivener et al. 2013); PROMs in hip and knee arthroplasty (Alviar et al. 2011), and performance based measures in hip and knee osteoarthritis (Dobson et al. 2012). It is important to note that while the methodological quality of some of the studies were marked down according to the criteria (Terwee et al. 2007), this may have been due to an omission by the authors when reporting the study, rather than in the conduct of the study. For example handling of missing data and the independent administration of tests may have been methodologically sound, but not reported in the literature. Improvements in standardised reporting of such studies was a key aim for the COSMIN group when implementing the checklist (Mokkink et al. 2010c).

The original LCI was studied most often, and was the only outcome measure which showed strong evidence (as defined by Terwee et al 2007 based on van Tulder et al 2003), for any of the measurement properties. Four studies presented results, either on the LCI section of the PPA (Miller et al 2001) or on a stand-alone version, (Gauthier-Gagnon et al 1998, Rushton et al 2002 and Franchignoni et al 2004) which is identical to the PPA section. Strong evidence was presented for its internal consistency and structural validity (Gauthier-Gagnon et al 1998, Miller et al 2001 and Franchignoni et al 2004). On this basis, clinicians and researchers have reasonable confidence for the validity of the LCI i.e. they know that is measuring the functional mobility of their patient. There was positive evidence from one study (Miller et al 2001) of fair methodological quality but with no comment on the measurement error, there is limited evidence for the variability of the LCI. A second study did present positive evidence of reliability and measurement error for the LCI, however the testing was done on consecutive days during a rehabilitation period; and due to a small sample size (n=37) the evidence was recorded as unknown (Franchignoni et al 2004). Clinicians should therefore be less confident about the

reliability and what constitutes a real change in scores for the LCI. The effect sizes reported by Franchignoni et al (2004) and Rushton et al (2002) were large, however with only effect size reported the evidence was deemed unknown and clinicians should use the LCI with caution, if they wish to measure the effectiveness of an intervention.

Franchignoni et al presented results on modified versions of the LCI in 2004 (LCI-5) and 2007 (LCI-5 and LCI-10-4) which helped to reduce the ceiling effect noted in the original version by increasing the number of response levels from 3 to 4 in the LCI-5. Little effect was noted in the Rasch analysis when the authors removed four items in the LCI-10-4 (Franchignoni et al 2007). These are the only studies, to date, that have presented psychometric results on these modified versions and it is interesting to note that while clinicians reported using the LCI-5, in addition to the original LCI, in the survey reported in Chapter 3, none reported using the LCI-10-4.

When measuring the impact of an intervention on the physical function of an amputee, e.g. a therapy programme or prosthetic component, it is essential that information on both the reliability (measurement error) and responsiveness is known, on any outcome measure being used (Streiner et al. 2014). Reliability, or consistency, of the results obtained from a test re-test design is expressed in correlations (ICC or Kappa). In addition, errors (random or systematic) that occur during completion of the test(s) will also affect the consistency of the results and are quantified by the measurement error value. The amount of measurement error will determine how reliable the outcome measure is considered to be. As it is presented in the same units of the outcome measure, the measurement error provides a meaningful guide for assessing confidence in the measured result (Stratford 2004). Knowledge of the quantity of error inherent in an outcome measure is therefore critical when detecting changes in clinical conditions with it.

While there were good quality studies that reported reliability results for many outcome measures in this systematic review, the evidence reported for measurement error on the same outcome measure was either not reported or unknown due to lack of comment regarding the clinical relevance of the level of agreement. Positive measurement error results were reported for the ABC scale (Miller et al 2003 and Sakakibari et al 2011), the L-test (Deathe et al 2005), the FAI

(Miller et al 2004) and RMI (Ryall et al 2003) but the evidence for these measures was considered limited due to low ("fair") methodological quality of the studies.

All studies reporting on responsiveness for any outcome measure were of fair quality only, and the results presented were considered "indeterminate" because of the author's choice of analysis. The magnitude of the change was most often reported using effect size. However, while effect size is an appropriate analysis tool to detect and quantify changes in the health status of a person, it *alone* is not considered an appropriate method of reporting the responsiveness of an outcome measure (Mokkink et al. 2010c, Terwee et al. 2003, Revicki et al. 2008). A triangulation of methods is recommended to establish the significance, magnitude and clinical relevance of any changes detected (Husted et al. 2000, Revicki et al. 2008, Beaton 2001).

The ongoing debate about responsiveness and how to measure it, leave some authors uncomfortable that traditional methods of defining and measuring it, are being challenged (Angst 2011). However, assuming that responsiveness is considered an aspect of (longitudinal) validity then it is reasonable to suggest that the outcome measure under study should be assessed by administering it at least twice during a longitudinal study. This was the case for all studies reporting responsiveness. A second outcome measure, preferably a gold standard, should also be administered at the same time to confirm that a change in the patient's condition, in the construct of concern, has taken place. A longitudinal study with a single outcome measure will not be able to determine both; the responsiveness of an instrument, as well as demonstrate the magnitude of the change in health status using the same instrument. The choice of analysis, in all studies reporting responsiveness, reduced the methodological quality to fair. The choice of analysis also affected the strength of the evidence and if it was considered inadequate, i.e. no comparator outcome measure or the use of effect size alone, it was assessed as unknown. Consequently there is an absence of evidence on responsiveness for all of the outcome measures included in this review.

4.5.2 Comments on quality issues

4.5.2.1 Demographics / aetiology

There was a variety of amputation levels and aetiologies, together with small sample sizes in many of the studies. This did not provide opportunity for comment or conclusion from authors, on convergent or divergent validity. In some studies, the aetiology of the sample was restricted e.g. all vascular or non-vascular participants (Panesar et al 2001, Schoppen et al 1999 and Hagberg et al 2010), which reduced the generalisability of the results to the wider population or pwLLA. The confounding effect of age on physical function in amputees (Sansam et al 2009) gives cause for concern with regard to the interpretability of some study results. Schoppen et al (1999) recruited only elderly subjects into their study. In contrast, a relatively young mean age (54 years) combined with a high prevalence of traumatic amputees, also reduced the generalisability of the results in the study on the LCI-5 and LCI-10-4 by Franchignoni et al (2007).

4.5.2.2 Study design

Many different study-designs were represented in this review. Test-rest methodology was used most often to measure reliability. A lower methodological quality rating was given when there was an inadequate explanation of the time interval between repeated measures; or when it was unclear whether participants were deemed to be stable (Rushton et al 2001, Franchignoni et al 2004). In contrast, data collected at various time-points through the amputee's rehabilitation journey, detected expected changes and allowed responsiveness indices to be calculated. However, when details of the interim period between tests were not given, a lower rating was awarded (Brooks et al 2001, Rushton et al 2002).

It is unusual to present details of both reliability and responsiveness using the same population sample during the same time period, though Ryall et al (2003) did so when reporting on the SIGAM, without detailing if there were different study periods. A lack of detail on how the overall effect size was calculated for different follow-up periods, and *a priori* hypotheses that had not been fully described contributed to the "fair" methodological quality rating for this study.

4.5.2.3 Sample size

A small sample size is considered a weakness when assessing both methodological quality and level of evidence. The COSMIN methodological checklist was developed for the assessment of HR-PRO questionnaires, which are completed by many respondents, and the thresholds for poor, fair, good and excellent in the checklist do appear stringent, especially with regard to sample sizes. It is more usual to find larger samples when testing HR-PRO questionnaires. However, it is prudent to consider the effect of different statistical analyses and sample size when assessing the level of the evidence. To avoid penalising studies with small sample sizes when assessing methodological quality and strength of evidence, sample size was considered only when the strength of evidence was being assessed. However, for many studies positive ("+") results for strength of the measurement property were reduced to indeterminate ("?") due to sample sizes being less than 50, according to the criteria. This was most noticeable in the reliability and measurement error categories where all studies had the strength of evidence level reduced because a sample size was < 50 (Table 4.4).

The difficulty of recruiting large participant numbers has been reported in other rehabilitation studies (Dobson et al. 2012, Scrivener et al. 2013), however the incidence of pwLLAs is lower than people undergoing hip or knee arthroplasty or stroke survivors. This low incidence results in a very challenging recruitment environment, especially considering that only 40% of amputees are fitted with a prosthesis (Scott et al. 2016).

If study-design and methodology are comparable amongst studies, results from multiple studies may be combined, thus improving the sample size. However, this was not possible in the current review as there were insufficient similarities between the studies.

4.5.2.4 Missing data

Two items, in each of the measurement property categories (except content validity) in the COSMIN checklist, refer to missing data. They are: "Was the percentage of missing items given?" and, "Was there a description of how missing data was handled?". Discussions took place between the two reviewers to ensure agreement of what was considered a missing "item" in each study i.e. whether it was an item of

a questionnaire or data from a whole visit, or both. It was further agreed that when a full data set was not presented, e.g. when only means and standard deviations or, only total scores for questionnaires were presented, it was not possible to deduce if all data had been presented for all participants. In nearly two thirds (22/36) of all studies reviewed, percentages of missing data could not be deduced and, there were no explanations about how any missing data was handled. This failure to address the two questions adequately resulted in a "fair" rating for methodological quality rating for all these studies.

Imputation analysis to account for missing data (both assessments and questionnaire items), is now being advocated. This is designed to remove unintended bias by leaving out any participants' data (Schafer and Graham 2002). However, no studies included in this review, mentioned its use.

4.5.3 Gaps in published evidence

In the 1990's and early 2000's, there was a push by professional bodies and networks to recommend the use of outcome measures in clinical practice (Hammond 2000, Jette et al. 2009, Abrams et al. 2006, Enderby and Kew 1995). The importance of reliability and validity was stressed and many new outcome measures were developed around this time, with work commencing on the content and structural validity of the measures. Studies then continued investigating both intra- and inter-rater reliability, with the outcome measures then being validated within different population groups and sub-groups. With the tightening of health budgets across the world, and the requirement to know the "value for money" of these interventions, there has been a rise in the interest of the responsiveness of any outcome measures in use.

The most recent studies investigating reliability of outcome measures of physical function used with pwLLAs were published in 2011 and 2013 in studies by Resnik et al (2011) and Wong et al (2013). However, before this, there had been a gap of nearly ten years in which no such studies were published. There had been a similar pattern of publication for validity studies with two studies published in 2013 (Wong et al and Major et al), but before this, there had been a gap of at least six years back to a series of studies that were published for the PEQ and LCI by

Franchignoni et al. At the time of searching for this review, only one responsiveness study had been published since 2004 (Kohler et al 2011).

Little has been published on the "clinical" responsiveness, i.e. the ability of an outcome measure to detect changes that are considered important either to the patient or clinician. Since this review, there has only been one study published that has reported clinical responsiveness of an outcome measure used with pwLLAs. The minimal clinically important difference (MCID) value was reported for the L-test when used with pwLLAs (n=33) (Rushton et al. 2014). The MCID was calculated using Receiver Operator Curve (ROC) analysis, which is deemed an adequate analysis by COSMIN. The methodological quality of the study was considered only fair because a description of the intervening period was not available. The level of evidence was negative because the ROC was < 0.7, however the final level of evidence was deemed "unknown", as the sample size was less than 50.

This increased interest in using outcome measures to measure the many aspects of prosthetic rehabilitation, has arisen from the increased financial accountability, not only within prosthetics but also in the wider healthcare field. It is therefore essential that clinicians and researchers fully engage with choosing and implementing the correct outcome measure, as well as the correct interpretation of their results. Demonstrating changes with individual patients is obviously important to the individual clinician and his/her patient, but demonstrating effectiveness and efficiency of whole services is also crucial.

4.6 Limitations

Several limitations have been highlighted throughout the production of this systematic review.

With additional titles found while examining the reference lists of included titles, a concern was raised that the search terms were not broad enough to cover all the inclusion criteria. It was not clear on examination of the additional titles, which criteria, i.e. population, intervention, condition or outcome, was compromised. However, the use of the published search strategy by Terwee et al (2009) to check for any missed psychometric measurement-property terms, did help to allay some

concerns. Terms related to prosthetics had not been used in the original search because it was thought they may widen the search too broadly. A repeat search with the addition of prosth* within the population search terms, confirmed that this did not provide a more focused population coverage with 3,582 titles returned. However there were four which were of interest to this review. Two presented on the Orthotics and Prosthetics Users' Survey (OPUS). The first detailed the development of the survey and presented internal consistency results following a Rasch Analysis (Heinemann et al. 2003). Results demonstrated good reliability with Person Separation Index (0.94) and Item Separation Index (0.98) for the Lower Limb Functional Measure section of the OPUS on 37 lower limb prosthetic users. However no further analysis was presented and, only a fair methodological quality rating was achieved due to the lack of information on how missing data was handled. A second study presenting validity evidence for a modified version of the OPUS also achieved a fair rating because of the missing data information (Jarl et al. 2012). A Rasch Analysis was again undertaken, following the addition of eight new items at the higher activity end of the scale, this time with only ten lower limb prosthetic users,. The results again demonstrated good reliability, but there was a ceiling effect seen. No further analysis was presented and, when the results from both studies are combined the evidence for this outcome measure would be rated unknown.

Reliability of the 6MWT in trans-tibial amputees was also presented in a study from India (Lahiri and Ghosh-Das 2012). However despite presenting good ICC values (0.79-0.88) for intra and inter-rater reliability no agreement data was presented and there were only 21 participants in the study, therefore this study does not alter the overall strength of evidence for the 6MWT presented earlier.

A study presenting initial reliability and validity results for the Orthotics and Prosthetics National Office Outcomes Tool (OPOT) was also found, though this tool primarily focused on measuring health-related quality of life and patient satisfaction and would not have been included in the review.

Many studies were excluded because their focus had been on evaluating the performance of either prosthetic componentry or instrumentation rather than that of the pwLLA, e.g. The AMPro and Transfemoral Predictor. While it is accepted that

these outcome measures do assess physical function of a pwLLA, predicting an amputee's "readiness" for ambulating with a prosthesis may not mean that the amputee completes the functional action. Validation studies of translated versions of outcome measures were also excluded as it was assumed that only one measurement property (cross-cultural validity) in the COSMIN checklist would be presented. It is possible that other psychometric measurement properties may have been presented within these studies however, the results are not presented here.

It is also recognised that there is a large psychosocial impact on the pwLLA following amputation. However, the degree to which factors such as self-efficacy and self-confidence influence the results of the self-reported measures was considered outside the scope of this thesis. In addition, studies investigating outcome measures used exclusively for measuring quality of life were excluded.

Finally, it is recognised that since the time the review was conducted in December 2013 there may have been some further studies published that would be of interest. Therefore a repeat search was done using the same basic search terms and only three studies were found that met the original criteria. Rushton et al (2014) presented MCID results for the L-test in a pilot study. Reid et al (2015) compared the validity of the 2MWT with the 6MWT and Franchignoni et al (2015) presented internal consistency results on a new questionnaire (Prosthetic Mobility Questionnaire) that is based on the PEQ.

4.7 Conclusions

There is a lack of high quality studies reporting the psychometric properties of physical function outcome measures used with lower limb amputees. The ratings of the methodological quality of some studies were reduced because of weak description within the manuscripts e.g. how missing data was handled. Poor choices for statistical analysis, especially with regard to responsiveness, together with small sample sizes of < 50, reduced the strength of evidence presented.

The LCI-5 has the strongest evidence to recommend its use for measuring functional mobility, however only for internal consistency and structural validity.

Nevertheless, there is a lack of corroborating data (i.e. more than one study) for many other outcome measures.

The lack of studies reporting on responsiveness, in particular the absence of MCID values, limits the number of outcome measures that can be recommended to investigate the effectiveness of any intervention on this population.

Confidence with which clinicians and researchers might make an evidence-informed selection of outcome measures for use with patients who have undergone a lower limb amputation is reduced, due to the limited availability of robust psychometric data that has been identified by this review.

4.8 Future work

Further evidence is required, especially in relation to the reporting of measurement error, responsiveness and MCID values, to improve the utility and clinical applicability of physical function outcome measures used with pwLLAs.

Standardised reporting of studies investigating psychometric properties of any outcome measure would enhance their methodological quality and strengthen the evidence presented. Collaborative work is also needed alongside standardised reporting to facilitate meta-analysis of data and reduce the impact of small sample sizes in this population.

Clear guidance on the use of statistical analyses should be made widely available with agreed protocols for their use to improve the reporting of measurement properties.

The impact of psychosocial factors following an amputation and their influence on the results of self-reported outcome measures should also be investigated.

5 Chapter 5 Repeatability Study (study III)

5.1 Purpose of chapter

The purpose of this chapter is to describe the methodology, and to present and discuss the results, of a study that was performed to establish reliability and agreement parameters of commonly used outcome measures in clinical practice that measure physical function during prosthetic rehabilitation. This study was carried out to answer Research Question 5.

Research Q5 What are the reliability and measurement error parameters of outcome measures most regularly used by AHPs for the assessment of physical function during prosthetic rehabilitation?

5.2 Background

A survey of AHPs, carried out in 2013 showed that the five most regularly used outcome measures in prosthetic rehabilitation for pwLLAs (adults) were: the SIGAM mobility grades; the LCI together with the modified version (LCI-5); a TWT of either two or six minutes duration; the TUG and SCS, see Table 5.1. Full details of the survey and the results were presented in Chapter 3.

Table 5.1 Outcome measures most regularly used in prosthetic rehabilitation

Name	**Type of Outcome Measure**	**Main construct being measured**
SIGAM	PROM	Functional mobility
LCI / LCI-5	Patient-reported outcome measure	Functional mobility
Timed Walk Test	Observed	Mobility endurance
TUG	Observed	Mobility and balance
SCS	PROM	Socket fit and comfort

The most regularly used outcome measures in prosthetic rehabilitation will be used within this study. Four of the five outcome measures listed are all concerned with measuring aspects of physical function. The SCS the final outcome measure listed in the top five is also a PROM. Although not measuring physical function per se, the SCS records the patient's perception of the fit and comfort of their socket which greatly influences their physical function (Gailey et al. 2008).

It was decided to also include a measure of global health in the study. The EuroQol standardised five dimension and five level instrument (EQ-5D-5L^{TM}) is recommended for use by the Chartered Society of Physiotherapists (currently only with musculoskeletal patients). The EQ-5D-5L is a measure of the global health of the respondent that can be applied across a wide range of health conditions and treatments to provide a simple descriptive profile and a single index value for health status (Rabin and Charro 2001). It is often used to describe the health of a population and the results can also be used to evaluate the clinical and economic status of health care, in particular the cost-effectiveness of certain interventions (Brazier et al. 1999). Knowledge of its test re-test reliability in the population of pwLLAs is unknown and therefore, its inclusion in this study is considered to be of value, both to clinicians and researchers interested in prosthetic rehabilitation.

Chapter 4 reported on a systematic review that investigated the published evidence on the psychometric properties for outcome measures measuring physical function of pwLLA prosthetic limb-wearers. When considering reliability and measurement error parameters in particular, it was found that 27 studies had reported inter-rater and/or intra-rater reliability for 30 outcome measures. However, only 13 out of these 27 studies had reported both reliability and measurement error results. Of the 15 studies reporting measurement error values, several different methods and terminology were used. Six studies had reported Bland and Altman plots and LoA; five presented SEM, while four presented MDC or equivalent (i.e. the smallest detectable difference (SDD), the smallest detectable change (SDC), the smallest real difference (SRD)) for a variety of outcome measures.

While the SEM estimates the standard error in a set of repeated scores the MDC relates to the smallest score that is beyond the measurement error of an outcome measure and therefore likely to represent a person's "true" score. Both are

meaningful for clinical practice, as they are presented in the unit of the outcome measure and the MDC can be derived from the SEM (Stratford 2004), represented by the following formula: MDC = 1.96 * √ 2 * SEM, where the 1.96 derives from the 95% confidence interval, and √ 2 is included because two measurements are involved in measuring change (Beckerman et al. 2001).

Previously published results for reliability and measurement error parameters on all the outcome measures included in this present study are presented in Table 5.2. The quality ratings and levels of evidence were obtained from the systematic review carried out and reported in Chapter 4. The reasons for the indeterminate level of evidence found in each study are given, in parenthesis, in the extreme right column of the table. The methodological quality ratings for each study are also given in the latter column, together with reasons for the rating. See Chapter 4 for full details of the quality review processes used to obtain these results.

Table 5.2 Published reliability and measurement parameters

Outcome Measure	**Study**	**Reliability parameters presented**	**Measurement error parameters presented**	**Level of Evidence / Methodological Quality (reason for ratings)**
SIGAM	Ryall et al (2003b)	Test re-test: ICC=0.79 Inter-rater: Kappa coefficient range = 0.86	None presented	Indeterminate ? (sample size) / Fair (details of how missing data was handled not given)
LCI / LCI-5 / PPA (LCI section)	Franchignoni et al (2004)	Test re-test: ICC =0.984 for both LCI and LCI-5	Bland Altman plot - no systematic variations for either LCI or LCI-5 version. LoA: -1.41-1.79 (LCI), -1.34-2.48 (LCI-5)	? (sample size) / Fair (conducted during rehab phase)
	Miller et al (2001)	Test re-test: ICC=0.88 for PPA-LCI	SEM = 2.7 for PPA-LCI	? (no comment on Limits of Agreement) / Fair (details of how missing data was handled not given)
TUG	Resnik et al (2011)	Test re-test: ICC= 0.88	MDC = 3.6s	? (sample size) / Good (administrations assumed to be independent, not confirmed)
	Schoppen et al (1999)	Spearman'st: Inter-rater=0.96 Intra-rater=0.93	None presented	? (sample size) / Fair (Spearman's cc only presented, no ICC)

2MWT	Brooks et al (2002)	Inter and intra-rater: ICC>0.98	None presented	? (sample size) / Fair (details of how missing data was handled not given)
	Resnik et al (2011)	Test re-test: ICC=0.83	MDC = 34.3m	? (sample size) / Good (test administrations assumed to be independent, not confirmed)
6MWT	Lin & Bose (2008)	Test re-test ICC = 0.94	Bland Altman plot - no systematic variation. LoA: -44.6-63.5 and -25.8-57.8 for two time periods.	? (sample size) / Fair (details of how missing data was handled not given)
	Resnik et al (2011)	Test re-test ICC = 0.97	MDC = 45m	? (sample size) / Good (test administrations assumed to be independent, not confirmed)
SCS	Hanspal et al (2003)	Inter-rater: Kendall's tau=0.99	None presented	? (sample size) / Fair (Kendall's tau only presented, no ICC)
EQ-5D-Index or VAS	None published in this population			

The lack of published evidence for agreement and measurement error values for these outcome measures, justifies the aim of this study which is; to establish estimates of reliability and measurement error for outcome measures regularly used by AHPs across the UK during prosthetic rehabilitation.

5.3 Methods

5.3.1 Participants

Adult (18 years or older) single lower limb amputees, at either TT or TF level, were eligible for inclusion in the study. They had to have been using a prosthesis for at least one year and wear it for at least 8 hours per day. With regard to their functional abilities, they were asked to declare themselves as an active outdoor limb-user, i.e. confirm that they had the ability to walk outdoors over different surfaces. In addition, as participants considered themselves in a stable state they were asked to report any issues, medical or otherwise, that might impact on their functional status during the period of the study.

Participants were excluded from participating in the study if they: had a history of any recent (within 3 months) prosthetic component changes or physiotherapy treatment; had any co-morbidities that prevented them undertaking the physical activities involved in the outcome measures or; if they had poor cognition or an insufficient comprehension of the English language that prevented them from understanding the written questionnaires. These exclusions were discussed with participants or with their clinical team as appropriate depending on the entry route to their participation, with the researcher making the final decision on who was included.

5.3.1.1 Demography and aetiology of study population

This was a convenience sample of self-selected amputees and twelve patients were recruited into the study. Table 5.3 presents the demographic and basic aetiological data for all participants.

Table 5.3 Demography and basic aetiology of participants

Age - yrs.	Mean (SD) Range	61 (15) 40-83	
	Median [IQR]	63 [49-73]	
Gender		n	%
	Male	10	83
	Female	2	17
Level	Trans-tibial (TT)	9	75
	Trans-femoral (TF)	3	25
Side	Right	7	58
	Left	5	42
Main Cause of amputation	Peripheral Arterial Disease (PAD)	1	8
	Diabetes	2	17
	Trauma	5	41
	Other	4	34
Number of co-morbidities	≤ Three	11	91
	≥ Four	1	9

The mean age was 61 years which is slightly younger than the mean reported for the amputee population in Scotland. The SPARG report, which is published annually presents data on every lower limb amputation in Scotland. The latest published report for the year 2013, reported a mean age of 67 years for adult lower limb amputees (Scott et al 2016). The male to female ratio in this study was slightly higher than that reported in the SPARG report however the TT to TF amputation ratio presented in Table 5.3 was similar to the SPARG numbers of limb-fitted amputees for 2013. The main cause of amputation was trauma with 41% or other (34%), with only 25% being peripheral arterial disease (PAD) or diabetes. The participants in this study do not appear to have a typical aetiological distribution for a group of pwLLA, compared to the percentages of referrals to prosthetic centres across the UK in 2011-12, published in the limbless statistics report (UNIPOD 2013). This report showed referrals for dysvascular reasons and diabetes accounted for 57% of the total, whereas referrals for amputation caused by trauma accounted for only 10%, and 33% of referrals were for other reasons.

5.3.2 Ethical opinion

Favourable ethical opinions were gained from both Queen Margaret University Ethics Committee and the NHS Ethics Committee (South East Scotland 02 committee) before commencing the study. NHS Lothian management approval was also sought and gained prior to conducting the study on NHS Lothian premises.

5.3.3 Recruitment process

The amputees were prospectively recruited into the study. A flow chart provided in Appendix 7 outlines the full recruitment process. Prosthetists, working at the Prosthetic Centre in Edinburgh, identified amputees who fitted the inclusion criteria described above from the centre's patient database. An invitation letter was sent to suitable patients outlining the study, together with a Study Information Pack (SIP) containing a 'Consent to Contact Form', Patient information Sheet (PIS) and a stamped addressed envelope. Participants were also recruited through an advertisement, see Appendix 8. The advertisements were placed in the fitting rooms of the prosthetic centre and uploaded to the website of the Edinburgh Limb Loss Association (ELLA) patient group. The chairperson of ELLA also sent the advertisement to all members via e-mail. Those amputees, who responded to the advertisement and approached their prosthetist, were also given the same SIP. Any interested members who saw the advertisement through ELLA contacted the researcher directly and were sent a SIP and the 'Consent to Contact Form' for further information.

The researcher made contact with any amputee who returned the Consent to Contact Form to answer any questions they had and, if they wished to take part, arrange a suitable date and time for the first test (TV1). The amputee was enrolled as a participant into the study at TV1 when written consent was obtained. At least 24 hours had elapsed after the amputee first received the SIP, or spoke to the researcher, before the TV1 occurred. This ensured that the amputee had adequate time to consider all the information given to them.

5.3.4 Study protocol

All study visits took place in the main Physiotherapy Department at Astley Ainslie Hospital, Edinburgh.

Baseline measurements for all outcome measures were taken at TV1, as well as basic demographic data recording age, cause of amputation, level of amputation and any relevant concomitant medical history.

Test Visit 2 (TV2) was planned exactly 7 days after TV1, when the outcome measures were repeated. See Appendix 9 for a full study schedule.

Every effort was made to complete the outcome measures at the same time of day for each participant's test visit. The order of the measurements was randomly assigned for each patient, at both visits, and Table 5.4 shows details of each of the study visits.

Table 5.4 Study visit plan.

Visit	What	When
Recruitment Visit / Telephone call (as required)	Explanation of Study Protocol and provision of Information Sheet & Consent Form	As soon as participant identified
Test Visit 1 (TV1)	Enrolment of participant i.e. consent taken and baseline outcome measures carried out	Arranged to suit participant
Test Visit 2 (TV2)	Repeat outcome measures	7 days after TV1

5.3.5 Outcome measures

Two walking tests and four PROMs were used in this study. A brief description of each is given below. The TUG test was performed in the physiotherapy gym where 3m was measured out on the floor and the same chair was used each time. The 2MWT was carried out in a straight corridor that was 35m long and free from obstacles and distractions. Data collection sheets and a detailed explanation of the procedures followed for each outcome measure can be seen in Appendix 10.

5.3.5.1 Walking tests

5.3.5.1.1 Timed Up and Go (TUG)

The TUG is a test of the manoeuvres required for basic mobility (Podsiadlo and Richardson 1991) and involves standing up from a chair, walking 3m, turning around a cone and returning to the chair and sitting down. The time taken to complete the circuit is recorded. In this study the following instructions were given: “Sit with your back against the chair and your arms on the armrest. When I say GO please get up and using your walking aid, walk around the cone, then return to the chair and sit back down. I am going to be timing, but this is not a race, please go at a pace that’s comfortable and safe for you.” At both test visits participants undertook the TUG three times, with at least a minutes rest between each time. The first occasion was to familiarise them with what was required, and the fastest of the second and third occasions was recorded as the test time.

5.3.5.1.2 Timed Walk Test (2MWT)

The 2MWT is a test used to measure functional walking capacity related to endurance capability (Butland et al. 1982). Participants are instructed to walk around a pre-measured circuit, covering as much ground as possible in the allotted two minute time period. Rests may be taken during the test and the total distance walked is recorded. Average speed may also be calculated, if it is of interest. Long straight corridors with few or shallow turns are recommended, but any circuit is acceptable as long as the repeat test is carried out on the same circuit. In this study the circuit used was a quiet corridor 35 metres long with 180 degree turns at either end. The participants were asked to perform the 2MWT only once at each visit.

5.3.5.2 Patient reported questionnaires

5.3.5.2.1 Locomotor Capability Index modified version (LCI-5)

The Locomotor Capability Index (LCI-5) is a self-administered scale specifically designed for use with pwLLAs (Franchignoni et al. 2007b). There are 14 questions about locomotor activities with 5 levels (0-4) available to score each question. The scores from each question are added together for a possible maximum score of 56 and a minimum of 0, with a higher score demonstrating better capability.

5.3.5.2.2 Specialist Interest Group in Amputee Medicine (SIGAM) Mobility Grades

The SIGAM Mobility Grades is a single-item mobility scale comprising of a self-reported questionnaire which has 21 yes/no items. When the answers to the questionnaire are applied to an algorithm a final grade of mobility is assigned from six clinical grades (A–F) (Ryall et al. 2003b). There are four sub-divisions for the C grade and three subdivisions for D that denote increasing levels of independence. The defined purpose of the SIGAM grades is to describe functional levels of mobility and the six grades describe a progression of increasing independence and functional locomotor ability from; limb abandoned/cosmetic only (A), through transfers only (B), walking less than 50m on even ground (C), walking greater than 50m on uneven ground (D), using an occasional walking aid (E) to normal or near normal walking (F). However, the scoring system takes into account increasing levels of walking support from a frame (a), two crutches or sticks (b), 1 crutch or stick (c) to none (d) for grade C and from (a) to (c) for grade D. The grades; A, B, Ca, Cb, Cc, Cd, Da, Db, Dc, E and F, were allocated numerical scores from 1-11 in this study.

5.3.5.2.3 Socket Comfort Score (SCS)

This score is a simple subjective measure of how comfortable the amputee feels the socket is at the time the score is taken (Hanspal et al. 2003). Amputees are asked a standard question: “On a 1 – 10 scale, if 1 represents the most uncomfortable socket fit you can imagine, and 10 represents the most comfortable socket fit, how would you score the comfort of the socket fit of your artificial limb at the moment?” Their response on the 10 point scale is then recorded.

5.3.5.2.4 EQ-5D-5L™

The EQ-5D-5L™ (EuroQol-5 dimension 5-level) tool is a self-administered instrument that measures health outcomes (Rabin and Charro 2001). It has five dimensions: mobility, self-care, usual activities, pain / discomfort and anxiety /depression. The respondent is asked to indicate their health status, in each dimension, on one of five levels: no problems, slight problems, moderate problems, severe problems, extreme problems or unable to do. A simple descriptive profile from the five dimensions is derived, which can be converted into a single index (EQ-5D-index) for health status, ranging from 1.000 (no problems in any dimension) to 0

(equivalent to being dead). The UK value set was used in the conversion calculator for this study. Also administered at the same time is the EQ (EuroQol) visual analogue scale (VAS). The EQ-VAS records the respondents self-rated health on a 20cm vertical VAS with end points labelled at 100 “the best health you can imagine” and at 0 “the worst health you can imagine.

5.3.6 Statistical analysis

The choice of analyses used was dictated by the type of data collected. Table 5.5 outlines the parameters of reliability and measurement error appropriate for continuous and ordinal data (de Vet et al. 2011).

Table 5.5 Choice of analysis methods

Measurement Property	Continuous data	Ordinal data
Reliability	ICC	ICC or weighted kappa
Measurement error / agreement	SEM or LoA	% agreement

Time and distance are considered continuous data containing equal ratios between each unit interval. With ordinal level data, the interval between each score or composite score may be variable because the total score is calculated from several individual items or is subjective to individual interpretation, e.g. VAS. Therefore, the 2MWT and TUG provide continuous (ratio) data; with data derived from the LCI-5, SIGAM, SCS and both the EQ-index score and EQ-VAS scale at the ordinal level.

Data from some ordinal scales have been considered continuous in previous studies with parametric descriptives, i.e. means, SDs, SEM and MDC values being presented (Resnik and Borgia 2011, Miller et al. 2001). Some authors (Jakobsson and Westergren 2005, Svensson 2001) have cautioned against this because of the uncertain and sometimes unknown ratio between data points. However, others are satisfied if the outcome measure, scale or questionnaire has been validated using Item Test Theory (ITT) techniques, such as Rasch Analysis (Franchignoni et al. 2007a, Franchignoni et al. 2007b, Gallagher et al. 2007, Gallagher et al. 2010) that

will confirm if the ratio between points on the scale is consistent throughout the scale (Andresen 2000, Tesio 2003, McPherson et al. 2015).

Scales, similar to the LCI-5 and SIGAM, where multiple items scores add up to a composite total score, e.g. the Roland Morris Questionnaire, the Western Ontario and McMaster Universities Arthritis Index (WOMAC), and the Berg Balance Scale have been considered continuous by many authors following a Rasch Analysis (Stratford et al. 1996a, Angst et al. 2001, Donoghue and Stokes 2009).

Franchignoni et al (2007) carried out a Rasch Analysis on the original LCI and the modified LCI-5 which confirmed excellent internal consistency and internal structural validity. Both the EQ-5D-index score and EQ-5D-VAS have also been regarded as continuous data and analysed as such (Walters and Brazier 2005). The argument for considering data collected in this study from SIGAM and SCS as continuous is not so strong because of the relatively small range of scores available, 11 and 10 respectively. However, a modified Rasch Analysis was completed for the SIGAM, on the shorter 1-6 grading system, by Ryall et al (2003), but to date there is an absence of either Rasch modelling or ITT analysis for the SCS. Therefore, data retrieved from the SIGAM, LCI-5, TUG, 2MWT and EQ-5D-5L in this study, was described and analysed as continuous data, while the SCS was considered as ordinal.

5.3.6.1 Missing data

When the data was inspected no data points were found to be missing in; individual items on questionnaires, total scores in the PROMs or, observed times and distance in the walking tests.

5.3.6.2 Summary descriptive statistics

Normal distribution of the data collected for 2MWT, TUG, LCI-5, EQ-5D-5L index score, EQ-VAS and SIGAM was assessed using Shapiro Wilks in order to determine whether the data was normally distributed.

Summary descriptive statistics were derived: means and standard deviations for parametric data; and medians and inter-quartiles for non-parametric and ordinal data.

5.3.6.3 Consistency analysis

Test re-test reliability was calculated for each outcome measure using ICC (model 2,1) (Shrout and Fleiss 1979). This two way ANOVA model was used because each of the participants was rated by a single rater and the results were to be generalised to other raters. For the SCS, the Cohen Kappa statistic was calculated (Scholtes et al. 2011).

5.3.6.4 Agreement analysis

Although not an appropriate measure of agreement in itself, a paired T-test or Wilcoxon ranked pairs test was performed on the outcome measures to confirm that none of the differences between the two repeated measurements were statistically significant.

Limits of agreement were calculated and visualised using a Bland Altman plot (Martin Bland and Altman 1986) for all outcome measures except SCS. Percentages of agreement and Kappa statistic were calculated for the SCS.

5.3.6.5 Measurement Error

Minimal detectable change values with a 95% confidence interval were calculated from the SEM. The SEM estimates the standard error in a set of repeated scores using the following equation: SEM = SD (baseline) X $\sqrt{}$ (1 - ICC).
The MDC value is calculated using the equation: MDC = 1.96 X SEM x $\sqrt{}$2.

The MDC% value was calculated by dividing the MDC by the average of the measurements (across both visits) and multiplying by 100.

5.3.6.6 Statistical software

All data was analysed using SPSS version 21.0.0.1, dated 2012, except percentage agreement which was performed in Excel ™

5.4 Results

Participants were assumed to be “stable” with no major issues with their prosthesis that required any prosthetic or therapy input at recruitment. While, no participant detailed any issues at TV1, one participant (participant P02R) reported an increase in the severity of his co-existing low back pain (declared at recruitment) during the period of the study, but did not consider that it had changed between the two visits. A second participant (P04R) mentioned that they were not feeling as well as they had at TV1 with regard to their concurrent gastro-intestinal medical problems. A third participant (P03R) reported having a slight head cold at TV2. These issues may have accounted for the changes noted for these participants, but other changes, both perceived and actual, were seen in all participants as can be seen in Table 5.6. Changes in perceived performances (recorded from the PROMS) and actual changes recorded from the observed outcome measures are noted here.

Table 5.6 Changes noted at second Test Visit

Participant	**Outcome Measures** (units)						
	SIGAM (grades 1-11)	**LCI-5** (points 1-56)	**TUG** (seconds secs)	**2MWT** (metres m)	**SCS** (points 1-10)	**EQ-5D-index** (points 0-1.00)	**EQ-5D-VAS** (points 0-100)
P01R	-1.0	1.0	-0.9	1.8	1.0	0	0.0
P02R	0.0	0.0	-0.9	10.9	0.0	0.408	0.0
P03R	0.0	1.0	0.5	-7.7	0.0	0	0.0
P04R	0.0	-3.0	-1.0	13.6	-4.0	-0.229	-35.0
P05R	0.0	-3.0	0.6	-13.3	0.0	-0.025	0.0
P06R	0.0	0.0	0.4	0.0	0.0	0	-10.0
P07R	0.0	0.0	0.4	14.1	0.0	0	10.0
P08R	0.0	-3.0	-1.6	0.4	1.0	0.017	0.0
P09R	0.0	0.0	-0.6	-4.4	1.0	0	0.0
P010R	0.0	0.0	0.6	-1.2	1.0	0	0.0
P011R	0.0	2.0	-1.8	7.1	-2.0	0.111	0.0
P012R	0.0	2.0	-0.6	16.0	0.0	0	-10.0
(improvement), (deterioration)							

5.4.1 Summary descriptive statistics

Shapiro Wilks test showed that the data were normally distributed for both test visits of the 2MWT and SCS, and TV2 for the SIGAM. All other data was not normally distributed (see Appendix 11).

Summary descriptive statistics are presented in Table 5.7, with means and standard deviations (SD) presented for parametric data, and medians, inter-quartile (IQ) ranges for non-parametric and ordinal data. Mean (SD) presented for 2MWT, Median [IR] presented for all other outcome measures. A full presentation of the data collected at both test visits can be found in Appendix 12.

Table 5.7 Summary descriptive statistics

Outcome Measure units (range)	**Test Visit 1 (TV1)**	**Test Visit 2 (TV2)**
SIGAM Grades (1-11)	10.0 [9.0 – 10.3]	10.0 [9.0 – 10.3]
LCI-5 Points (0-56)	53.0 [48.8 – 55.3]	52.5 [47.0 – 55.3]
TUG Time (secs)	10.1 [9.4 – 10.8]	10.1 [8.9 – 10.6]
2MWT Distance (m)	121.0 (26.9)	124.2 (24.9)
SCS Points (0-10)	9.0 [7.8 – 9.0]	9.0 [7.8 – 9.0]
EQ-5D Index Points (0-1.000)	0.802 [0.699 – 1.000]	0.923 [0.753 – 1.000]
EQ-5D VAS Points (0-100)	87.5 [80 – 95]	90 [80 – 95]

5.4.2 Inferential statistics

5.4.2.1 Consistency analysis

Table 5.8 presents ICC and kappa statistics on the test re-test values obtained at TV1 and TV2. Intra-class correlations of >0.9 were found for TUG, 2MWT, LCI-5 and SIGAM. Values of >0.75 presented for ICC, and other indices of reliability, indicate an excellent level of practical and clinical significance for test re-test reliability as summarised by Cichetti in 1994 from the work of Landis and Koch (1977), Shrout and Fleiss (1979), and Cichetti and Sparrow (1981) (Cicchetti 1994). Good reliability was found for both the EQ5D index scores and VAS with ICC values >0.60 and <0.74 (Cicchetti 1994). A Kappa statistic of 0.314 was presented for the SCS which represents fair reliability (Landis and Koch 1977).

Table 5.8 ICC (2,1) for all outcome measures except SCS, Kappa statistic

Outcome Measure	Intraclass Correlation Coefficients (95% CI)
SIGAM	0.968 (0.894 – 0.991)
LCI-5	0.972 (0.906 – 0.992)
TUG	0.991 (0.968 - 0.997)
2MWT	0.936 (0.794 – 0.981)
SCS	0.314
EQ-5D Index	0.728 (0.294 – 0.913)
EQ-5D VAS	0.704 (0.247 – 0.905)
ICC≥0.75 = excellent, ≥0.6 and <0.75 = good, <0.6 = fair	

5.4.2.2 Agreement analysis

Limits of agreement were calculated and visualised using a Bland Altman plot (Bland and Altman 1986) for all outcome measures except SCS. The mean difference was calculated by subtracting the mean of TV2 from the mean of TV1 and the LoA were calculated as 1.96 x SD of the mean differences. As can be seen by the mean differences presented in Table 5.9, there is a deviation away from zero for all the outcome measures. This demonstrates a slight bias towards the first visit with a worse score at the second visit for LCI-5, EQ5D-VAS and SIGAM outcome measures; and a slight bias towards the second visit for the TUG and 2MWT, and EQ5D-index measures.

Table 5.9 Mean difference between (TV2-TV1) and Limits of Agreement

Outcome Measure units (range)	**Mean difference (LoA)**	**Range of LoA 1.96 x SD diff**
SIGAM Grades (1-11)	0.1 (-0.5 – 0.7)	1.2
LCI-5 Points (0-56)	0.3 (-3.3 – 3.8)	7.1
TUG Time (secs)	0.4 (-1.3 – 2.1)	3.4
2MWT Distance (m)	-3.1 (-21.3 – 15.1)	36.4
EQ-5D Index Points (0-1.000)	0.024 (-0.304 – 0.256)	0.560
EQ-5D VAS Points (0-100)	4 (-18 – 26)	44

Examination of the Bland Altman plots below, presented in Figures 5.1 – 5.6, show the plotted mean differences in the context of the LoA. This allows comment on the magnitude of these limits and also whether any data points are lying outwith them.

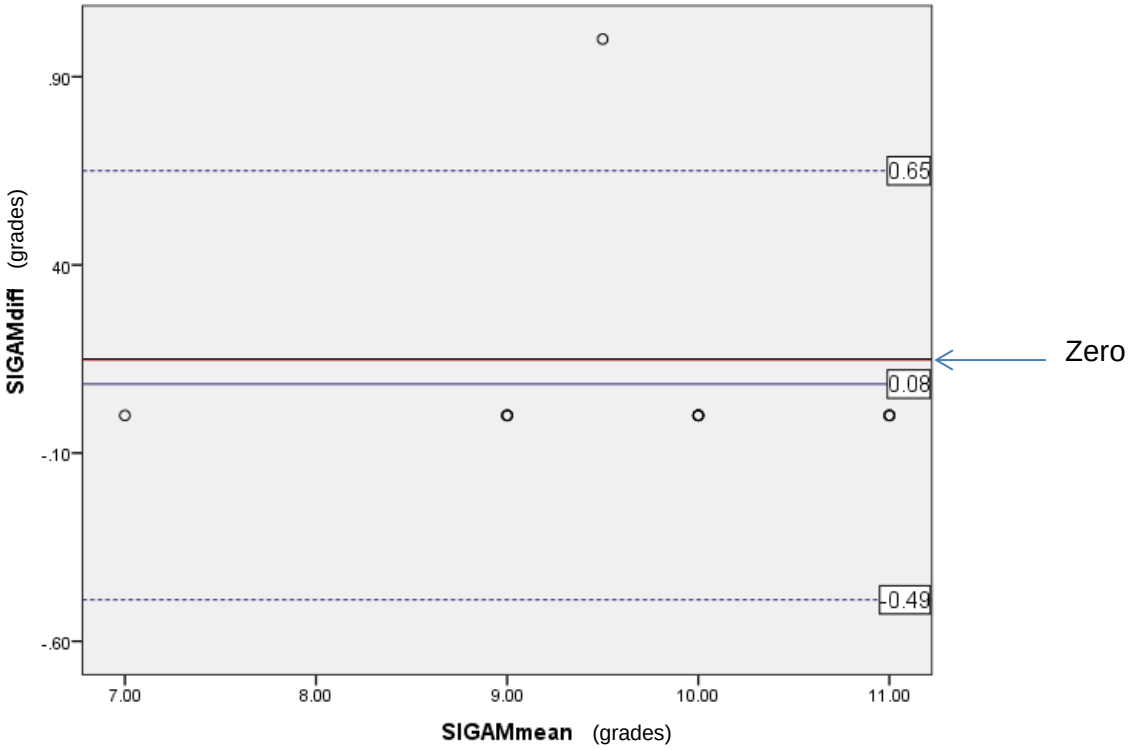


Figure 5-1 SIGAM Bland Altman plot

There was one participant who changed his / her SIGAM score by 1 grade between the first and second visit and this showed up as an outlier on the Bland Altman plot (Figure 5.1).

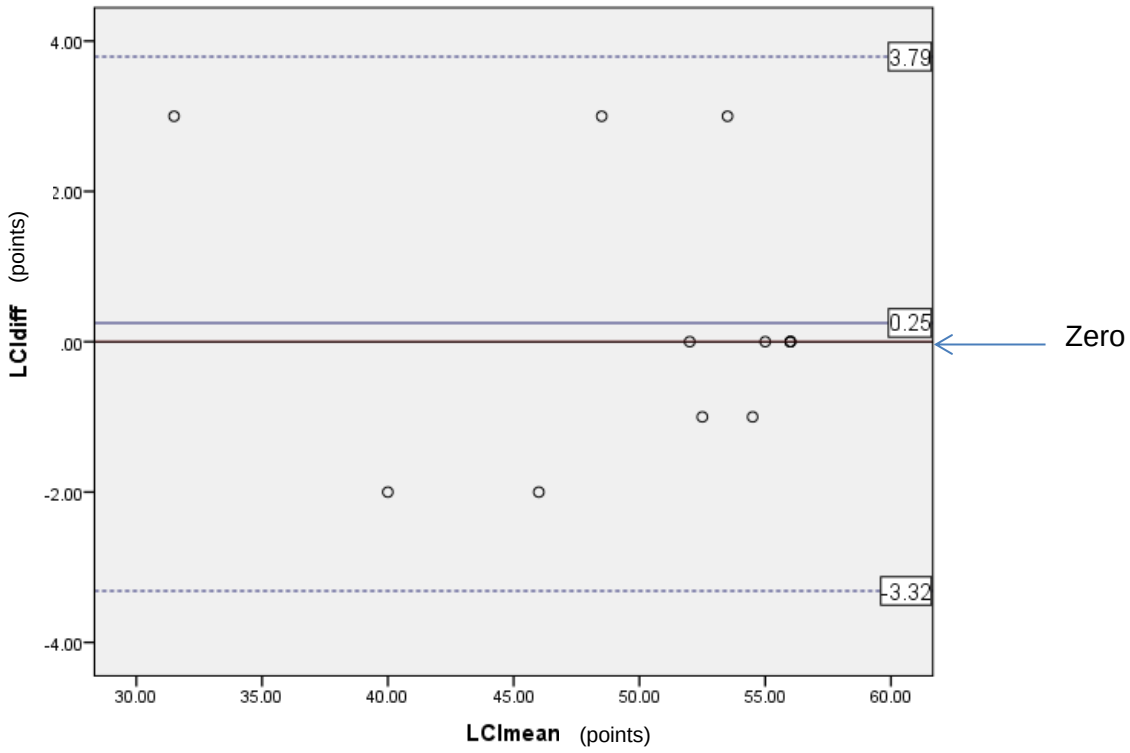


Figure 5-2 LCI-5 Bland Altman plot

All data points for the LCI-5 were within the limits calculated, but those participants scoring closest to the maximum (56) appeared to have a higher agreement between the two measurements compared to those with lower scores (Figure 5.2). The deviation from zero of the mean difference was minimal (0.25) and unlikely to be considered clinically significant as it represent a change of less than one score on the scale between the two visits. However, LoA for the LCI-5 of 7 points may be considered wide when interpreting variability in the clinical performance, but five participants had a difference of 2 or more points between the two visits.

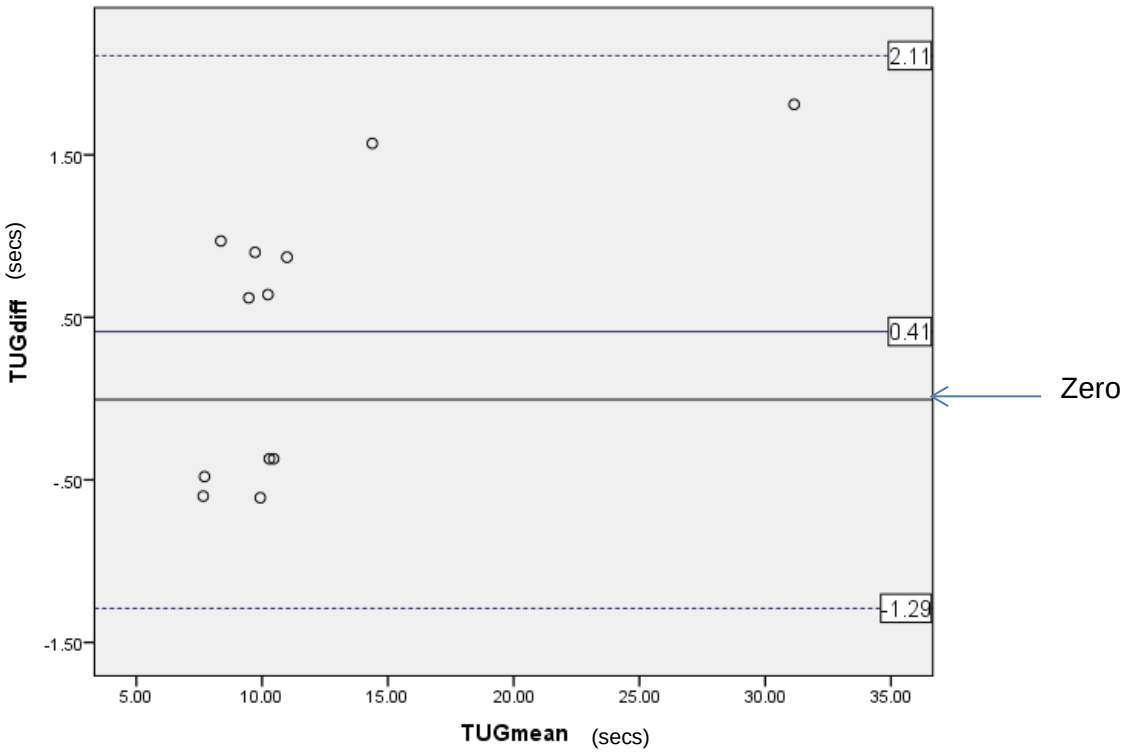


Figure 5-3 TUG Bland Altman plot

All data points are within the LoA for the TUG (Figure 5.3). The two slowest participants were both quicker on the second visit and when their data were removed, the mean moved closer to zero, from 0.4 to 0.2s and the LoA reduced from 3.4secs to 2.7sec. When considering the clinical performance of a stable amputee, the width of the LoA of 3.4s (for all the participants) still represents a reasonable amount of variability for an individual pwLLA.

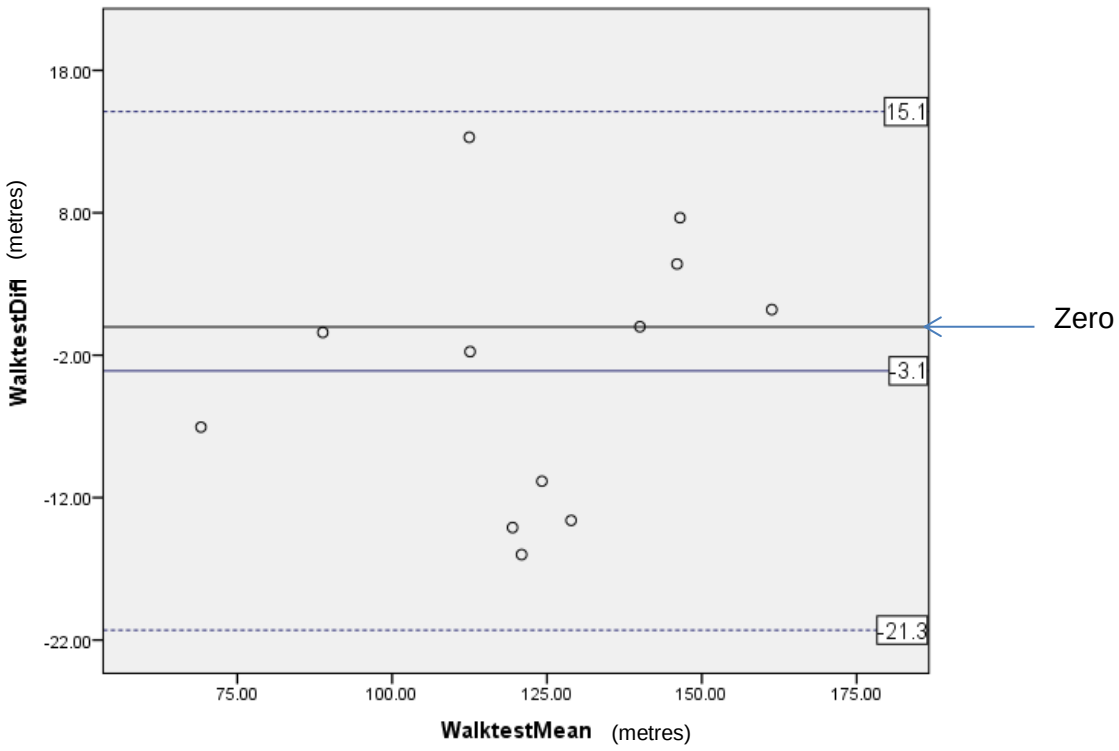


Figure 5-4 2MWT Bland Altman plot

All data points were within the limits in the plot for the 2MWT, and there was no proportional bias noted. The (negative) deviation from zero of the mean difference demonstrated a mean increase of 3m at the second visit, which is unlikely to be of clinical significance when taking in account the total distances recorded for this group of participants across each visit ranged between 74m and 160m. However, the LoA is wide (-21.3–15.1, i.e. 36.4 m), when considering the clinical performance of a stable amputee.

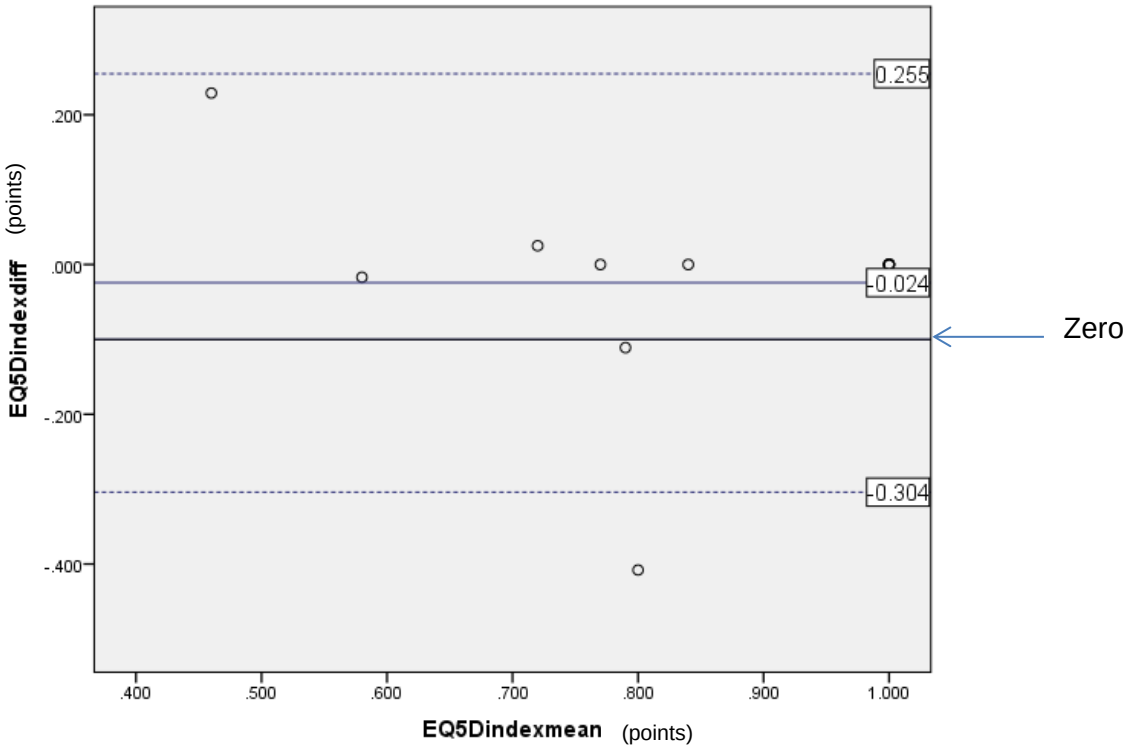


Figure 5-5 EQ-5D-index Bland Altman plot

There was one outlier outside the lower limit, lying close to the upper limit of agreement on the Bland Altman plot for the EQ-5D-index (Figure 5.5), and one that was just inside the upper limit. All the other data points were lying close to the mean (-0.024). With both outliers removed the mean moved closer to zero (-0.01) and the width of the LoA reduced from 0.559 to 0.146

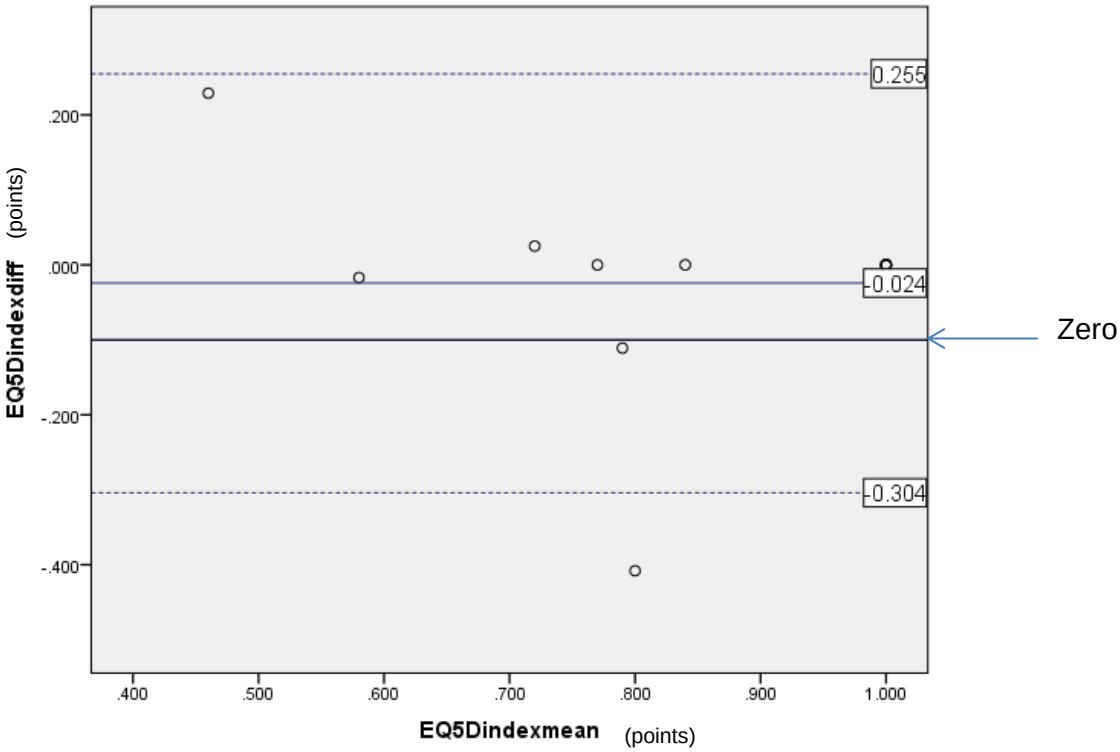


Figure 5-6 EQ-5D- VAS Bland Altman plot

There was one outlier outside the upper limit for the EQ-5D-VAS scores; whose mean score was 52.5 (see Figure 5.6). The remaining data points were clustered to the right of the plot closer to zero, demonstrating a trend towards a higher level of agreement as the mean scores increased.

The percentage agreement for SCS was reported as 50%. This was higher than the percentage for LCI-5, which at 42% was the lowest, and SIGAM demonstrated the highest percentage agreement at 92%, (see Table 5.10). The percentage agreement was not calculated for the continuous data collected for the TUG and 2MWT.

Table 5.10 Percentage agreement

Outcome Measure	Percentage Agreement between TV1 and TV2
SIGAM	92%
LCI-5	42%
SCS	40%
EQ-5D-index	58%
EQ-5D-VAS	67%

5.4.2.3 Measurement error

Table 5.11 presents the mean and SD for each visit together with the mean difference for each outcome measure. The information below is basically the same as the Bland Altman plots with the MDC being half of the width of the LoA.

Table 5.11 TV1 and TV2 means (SD), SEM, MDC and MDC%

Outcome measure (units)	TV1 mean (SD) [range]	TV2 mean (SD) [range]	(TV1-TV2) mean difference (SD) [range]	SEM	MDC	MDC %
SIGAM (grades)	9.8 (1.2) [7 – 11]	9.7 (1.2) [7 – 11]	0.1 (0.3) [0 – 1]	0.2	0.6	6
LCI-5 (points)	50.3 (7.8) [33.0 – 56.0]	50.0 (8.0) [30.0 – 56.0]	0.3 (1.8) [-2.0 – 3.0]	1.3	3.5	7
TUG (secs)	11.9 (8.0) [7.4 - 32.1]	11.5 (7.0) [7.9 - 30.2]	0.4 (1.0) [-0.6 – 1.8]	0.6	1.8	15
2MWT (m)	121.0 (28.1) [65.6 - 161.9]	124.1 (25.9) [72.7 - 160.7]	-3.1 (9.3) [-16.0 – 13.3]	6.8	18.9	15
SCS (points)	8.5 (1.1) [7 – 10]	8.3 (1.5) [5 – 10]	0.2 (1.5) [-1 – 4]	0.9	2.5	29
EQ-5D-index (points)	0.818 (0.186) [0.575 – 1]	0.842 (0.218) [0.348 – 1]	-0.020 (0.140) [-0.408 – 0.025]	0.09	0.26	31
EQ-5D-VAS (points)	87.5 (10.3) [70 – 100]	83.8 (18.2) [35 – 100]	3.8 (11.1) [-10 – 35]	5.4	15.5	18

The SEM and MDC values for the ordinal scale SCS were calculated using mean and SD values. As previously stated, this is not recommended for ordinal data and therefore the results should be viewed with caution. The MDC% values quantify the percentage change, with respect to the participant population mean values.

5.5 Discussion

The focus for this study was on estimating reliability and measurement error for those outcome measures of physical function that are used regularly during the prosthetic rehabilitation of pwLLAs. Values for consistency, agreement and measurement error, were all calculated using a test re-test study design with an intervening period of 7 days between tests.

This is the first study to present reliability indices for the EQ-5D-5L in a pwLLA population. It is also the first time that both consistency and agreement parameters (including comments on LoA) have been presented for the SIGAM, LCI-5, TUG and 2MWT in this population. The MDC values stated for the SIGAM, LCI-5, and the EQ-5D-5L are also the first to be reported in relation to pwLLAs for these outcome measures.

5.5.1 Consistency and agreement parameters

Reliability is considered the extent to which scores are the same when tests are repeated under the same conditions i.e. how consistent the scores are, and how close in agreement they are, when the tests are repeated. Consistency within repeated measurements is often reported by means of ICCs, as in this study, using repeated ANOVA statistics. However, correlations do not consider any systematic bias that may be present and the level of agreement between two test scores must also be taken into account, when commenting on the overall reliability of an outcome measure.

Intra-class Correlation Coefficients of >0.9 were calculated for the TUG, 2MWT and LCI-5, representing excellent consistency (Cicchetti 1994). This confirmed previous results seen for these measures: Schoppen et al (1999) reported Spearman's coefficients of 0.96 (inter-rater) and 0.93 (intra-rater) reliability of the TUG in 1999; ICC of 0.88 was reported for the 2MWT by Brooks et al (2002) and Franchignoni et al reported 0.98 (ICC) for the LCI-5 in 2004. Resnik et al (2011) reported good reliability (>0.7 and <0.9) for the TUG with ICCs of 0.88 and 0.83 for the 2MWT. The ICC value for the SIGAM, was also better in this current study (>0.9) than in the only other previous study (n= 62), which reported good reliability (ICC = 0.79) (Ryall et al 2003).

With ICC (95% CI) values all >0.70, the TUG, 2MWT, LCI-5 and SIGAM could all be considered reliable enough to be used clinically on the basis of their consistency results. A value of 0.70 is the minimum recommended level at which a test can be used with individuals, as opposed to groups in research (Streiner et al. 2014).

In addition to excellent consistency, the TUG and SIGAM also showed good levels of agreement with low bias and narrow LoA in the Bland Altman plots. Thus both outcome measures have shown good overall reliability for when used with pwLLAs. Previous Kappa coefficients, varying between 0.87 and 0.93, reported for the SIGAM in a previous study (Ryall et al 2003) confirmed good agreement of the scale when scored by patients and clinicians. However, no comments on LoA were recorded by Ryall et al (2003). With excellent consistency and good agreement, the SIGAM would appear to be of use to clinicians when evaluating functional mobility change in pwLLAs. There have been no reported comments on agreement parameters for the TUG in any previous studies with pwLLAs known to date; therefore this study presents the first of this kind. From these results the TUG would appear to be of use to clinicians and researchers when evaluating balance and mobility with pwLLAs.

The wide LoA seen for the 2MWT (36.4m) may have compromised the level of agreement, as this distance constitutes nearly one third of the mean distances recorded during each of the two test visits. Wide LoA for the 2MWT have been seen previously. A study presented normative values for the 2MWT from a sample of 1137 (age range 18-83), and the authors stated that the Bland Altman plot, calculated from the sample, did not reveal any systematic error (Bohannan et al 2015). While this was probably correct, it is debateable whether the wide LoA (84.9m) presented, represents an acceptable level of agreement in a stable sample. The authors did comment that the multiple centres in the trial were a possible source of random error, but this is not likely to impact on LoA. The wide age range of their participants was also cited as being partly responsible for the extensive variability reported. This may have been because slower walking speeds, which are associated with increasing age and increasing disability levels, have been shown to increase within-day variation (Feys et al. 2014). It is known that TF amputees walk slower than those amputated at the TT level (Genin et al. 2008). However, the Bland Altman plot for the 2MWT in this current study, did not show a trend towards

the slower walking speeds (as represented by the smaller total distances), despite the age range of 40 to 83. On the evidence from this study, it is would appear that the 2MWT is suitable for interpretation at the group level only with pwLLAs. The reporting of such wide LoA would suggest that the 2MWT may not be reliable for individual use in this population.

Seven points may be considered wide LoA for the LCI-5 (-3.3–3.8) when interpreting variability in the clinical performance of a pwLLA. The functional performance of a stable amputee is unlikely to decrease or even increase to the extent that there would be a change in score of more than 2 points on the LCI-5 in the space of one week, unless they started or stopped using a walking aid for some reason. Neither the researcher, nor the participant, had access to the previous week's results and therefore any change in scores had not been discussed or compared to their previous answers. On closer inspection of the answers given by the five participants who registered a change of more than 2 points between the two visits found that they stated that; they either required a walking aid when previously they had not, or vice versa. Though the reasons for this apparent change in dependence on a walking aid were not collected, it is assumed that there had been a change in confidence levels that may have affected their perceived locomotor abilities. Interestingly, two out of the three participants whose scores dropped by 3 points, indicating an increased dependence on walking aids or stand-by assistance, walked quicker in the TUG and further during the 2MWT. One of these participants also registered a 4 point improvement in the comfort of their socket on the SCS. Kendall's tau correlations between the change scores registered by the LCI-5 and the other outcome measures did not reveal any close relationships. None were higher than 0.313 (LCI-5 and EQ-5D-index). The relationship between perceived and observed outcome measures will be explored further in the overall thesis discussion in Chapter 7.

Limits of Agreement for the LCI-5 have been presented in only one previous study (Franchignoni et al 2004) and were narrower (-1.3-2.5) than those presented here. Despite the authors of that previous study stating that the LoA demonstrated "good reproducibility" for the LCI-5, it can be argued that those limits are still wider than may be expected when repeat measures are only one day apart as was the case in

that study. From the agreement evidence presented in the current study, it is therefore recommended that the LCI-5 be used for group level analysis only.

The EQ-5D-5L is concerned with general health overall with respect to five individual domains, and while not all five are directly related to physical function, the areas covering pain and discomfort, and anxiety and depression, will have an impact. How much this is on a day to day basis is not known in pwLLAs, but pain and anxiety were the health problems that the English public deemed as most important to them, as measured by the EQ-5D-5L (Devlin et al. 2016). It is difficult to comment whether the agreement limits reported are representative of reasonable variability in the clinical performance of the participants, as this is the first presentation of reliability for the EQ-5D-5L in a population of pwLLA. However, both the EQ-5D-index score and VAS registered good ICCs of between >0.60 and <0.74. It is therefore recommended that the EQ-5D-5L should be used at the group-level only with pwLLAs until further data is collected.

This is also the first study reporting agreement for the SCS. However with an ICC of 0.336, a Kappa statistic of 0.314 and only 50% agreement in the repeat scores, the SCS demonstrated poor overall reliability and may be limited for either group or individual use.

5.5.2 Measurement error

Measurement error may be reduced by performing repeated measurements and calculating average scores (de Vet et al. 2011 pg 107). This principle of reducing measurement error is also applied when outcome measures are investigated for research purposes, in groups of patients and repeat measures are undertaken. Higher ICC values are required to recommend its use with individual patients (de Vet et al. 2011 pg 142). The implication being that when results are obtained from a group in a research study, they are reduced by a factor dictated by the sample size. Consequently, any decisions in clinical practice taken for individual patients, should consider a MDC higher than that calculated from the group and greater changes must be detected to be sure their patient has made a true change (beyond random error) (de Vet et al. 2011 pg 244).

For the SIGAM, the MDC_{95} was calculated to be 0.56 and therefore a difference of at least 1 grade (when using the eleven grades) would be required to see a real difference in the level of an amputee's functional mobility.

With a MDC_{95} value of 3.5 for the LCI-5, this means that a difference of at least 4 points is required, to detect a real change in the functional mobility level of an amputee, on the 56 point scale. These are the first MDC values to be presented for the LCI-5 and SIGAM, in a population of pwLLA. With no previous measurement error results presented for either the LCI-5 or SIGAM comparisons with other studies are not possible. Miller et al (2001) did present SEM value for the original LCI, however the scoring system changed from a three level in the original scale to a five-level system in the LCI-5, so the values will not be comparable to this study.

The MDC_{95} value for the TUG was calculated to be 1.8secs for the 12 participants in this study, but clinicians should look for time changes greater than this to be sure that a change has occurred in their patient's condition with regard to their walking and balance abilities. This value represents 15% of the mean values over the two visits. Results from the current study showed that the MDC_{95} was 18.9m for the 2MWT, which also represents 15% of the mean distances covered by the group across the two visits. Both these MDC_{95} values and percentages are considerably smaller than those recorded in a similar population by Resnik et al (2011). They reported MDC_{90} values of 3.6s for the TUG (28%) and 34.3m (29%) for the 2MWT in a population of pwLLAs (Resnik et al 2011). There were two raters in the Resnik study with only one in this study, which may have reduced the random error and consequently contributed to the increased reliability recorded here. The higher proportion of TF amputees in the Resnik study sample, compared to the current study, may also have contributed to the increased variability seen in their results. As previously stated; TF amputees walk slower than those amputated at TT level (Genin et al. 2008) and slower walking speeds may increase day to day variability. The normative study by Bohannan et al (2015) calculated a value of 42.5m (MDC_{95}) from a subset of 157 participants (Bohannon et al. 2015), which although higher than the MDC reported by Resnik et al (2011), represented a lower percentage (25%) of the mean of the distances measured.

The MDC values observed, in this current study are; lower for the TUG and, within the ranges for the 2MWT, presented for other patient populations. Values for the TUG range from 2.9s in a chronic stroke population (Flansbjer et al. 2005) to 11s in long standing Parkinsons Disease (Steffen and Seney 2008); and for the 2MWT 13.4m in stroke survivors (Hiengkaew et al. 2012) to 22.9m in subjects with poliomyelitis (Stolwijk-Swuste et al. 2008). The MDC values presented in this study will add to the evidence for clinicians to use as reference values for their patients; and for researchers planning intervention studies.

A change score of at least 0.258 (MDC_{95}) on the EQ-5D index score and 15.5 on the EQ-5D VAS would be required to see a real difference in the global health of a pwLLA. While measurement error values have not been presented in a similar population, SEM values have been presented in a study of cancer sufferers. The SEM value for the EQ-5D-index scores was 0.11 for the whole group (n=534) and 0.12 for the smaller sub-set of lung cancer (n=50), both of which are similar to the SEM values recorded for this group of pwLLAs (0.10). As these are the first measurement error values presented for the EQ-5D-5L in a population of LLAs, they will provide reference values for researchers planning future studies.

Minimal detectable change values were not calculated for the SCS. This was because the SCS is an ordinal scale and, as such it was not considered appropriate to calculate SEM values using means and SDs.

5.6 Limitations

This study has the following limitations:

As this was a single researcher PhD study inter-rater reliability was not investigated.

The sample size was small and as such, the methodological quality of the study would be considered poor if assessed using the COSMIN quality criteria, and the quality of evidence would be deemed indeterminate. However, it could be argued that having a higher sample size may have resulted in lower agreement with the increased possibility of higher variances in the participants. There was reduced variability demonstrated in this study compared to other similar studies for the TUG

and 2MWT (Resnik et al 2011) and the SIGAM (Ryall et al 2003) that had higher sample sizes.

Any generalisation to the wider population of pwLLA must be made with caution, especially to vascular amputees with diabetes. The proportion of participants with an amputation due to a vascular or diabetic cause was smaller in this study, compared to published figures for all amputations (Scott et al. 2016), though they were closer to those published for amputees referred for limb-fitting (United National Institute for Prosthetics,& Orthotics Development 2013).

The MDC values have been estimated using a study group who have been walking on their prostheses for at least a year and generalisation of these values to those who may be in the early stages of their rehabilitation should be done with caution. However, it is still valuable to provide a guide to the error measurement for these outcome measures when used with pwLLAs.

In clinical practice, a change score is based on the difference between pre-treatment and post-treatment scores, and the treatment period may extend to several weeks for an amputee. When assessing test re-test reliability in a stable population, this would ideally be over the same time-period, however in practical terms this is not always possible. Shorter test re-test periods, such as one week in this study, are often used and the assumption is made that the measurement error detected during this shorter period will be applicable to the longer periods seen in longitudinal intervention studies and rehabilitation programmes.

5.7 Conclusions

From the results of this repeatability study carried out on a small group of established stable pwLLAs the following conclusions have been drawn:

Of all the outcome measures studied, only the SIGAM and TUG demonstrated excellent reliability for both consistency and agreement over repeated measures, when used with an established stable population of pwLLA. On this basis, they can both be recommended for use at the individual level.

A change of at least one grade is required on the SIGAM, when the eleven grades are used, to demonstrate a real change in the functional ability of a pwLLA.

A change of at least 1.8secs on repeat TUG measures is required to demonstrate a real change in the balance and mobility of a pwLLA.

Both the 2MWT and LCI-5 demonstrated excellent consistency but the limits of agreement were questionable and therefore cannot be recommended for use at the individual level. However, they may be considered reliable at the group level in this population. It is recommended that more data is collected on the 2MWT and the LCI-5, and further agreement analysis is performed, before reliability can be confirmed at the individual level for these measures when used with pwLLAs.

A change of at least 18.9m on repeat 2MWT measures is required to demonstrate a real change in the mobility and endurance of a pwLLA.

A change of at least 4 points is required on the LCI-5 to demonstrate a real change in the functional abilities of a pwLLA.

Good consistency was seen in both the EQ-5D index scores and VAS within repeated measures, but agreement was difficult to comment on. It is recommended that more data is collected on pwLLAs with the EQ-5D-5L, before its reliability can be confirmed in this population.

6 Chapter 6 Longitudinal Cohort Study (study IV)

6.1 Purpose of chapter

The purpose of this chapter is to describe the methodology, and to present and discuss the results of a study that was performed to establish the responsiveness of the most commonly used outcome measures in clinical practice measuring physical function during prosthetic rehabilitation. This study was carried out to answer Research Question 6.

Research Q6 What is the responsiveness of physical function outcome measures regularly used with lower limb amputees when assessing change during the rehabilitation period immediately following limb fitting?

6.2 Background

6.2.1 Study context

The setting for this study is the early prosthetic rehabilitation period, i.e. the period following the amputation surgery, when a pwLLA is fitted with his or her first prosthesis. During this time, initial gait training and therapy is provided. This early period also includes the immediate post-hospital discharge phase, when the pwLLA begins to adjust to his or her home circumstances with a prosthetic leg. As the patient recovers from major surgery, changes will be occurring in their clinical condition and capabilities. It is anticipated that improvements will be taking place with regards to their mobility and there will be increased participation in all activities of daily living, as they become accustomed to using a prosthesis for the first time. Clinicians are interested in measuring such progress, which can additionally be used to motivate their patients during these early stages of their rehabilitation. Any outcome measure used to evaluate the impact of an intervention, such as gait training, a muscle strengthening programme or provision of a prosthesis, must be sensitive to changes in the amputee's condition and physical abilities in whatever construct it purports to measures. The ability of an outcome measure to be sensitive to change is considered one of its psychometric properties, which is also

referred to as its responsiveness (Kirshner and Guyatt 1985, Guyatt et al. 1989, Husted et al. 2000).

Assessing the responsiveness of an outcome measure to changes in a patient's condition is not the same as measuring the response of a patient to an intervention. When studies report changes in the physical condition or abilities of a particular patient population, in a particular setting, the assumption is often that the outcome measures used in these studies are sensitive, or responsive to these changes (Beaton 2001). Although the same statistical methods and indices may be used when assessing the responsiveness of an outcome measure or when investigating patient responses; different hypotheses need to be tested in each case. In this study, it is assumed, based on clinical experience, that there will be changes in the physical function of the participants, as they recover from major surgery, undergo gait and strengthening training as well as being fitted with a prosthesis. Therefore, the focus of the study will be on calculating the responsiveness of the outcome measures, rather than on measuring the magnitude of the response to any of the interventions.

6.2.2 How is responsiveness of an outcome measure represented?

Responsiveness has previously been recognised as an aspect of validity, but more recently it is being considered separately (Husted et al. 2000, Beaton 2001, Mokkink et al. 2010b). The distinction between validity and responsiveness being that, validity refers to a single score whereas responsiveness considers a change score based on two measurements (Guyatt et al. 1987).

The aim of a recent review of the literature was to investigate the responsiveness of commonly used outcome measures that are used to measure physical recovery in the lower leg, ankle and foot. However, a lack of consistency in how responsiveness was defined and reported in the studies chosen, led the authors to call for further work, in order that a standard reporting measure for responsiveness could be recommended (Shultz and Olszewski 2013).

6.2.2.1 Distribution-based methods of calculating responsiveness

As detailed in Chapter 2, it is known that there are many indices that can be used to report the statistical significance of change scores as well as their direction and magnitude. Statistically significant differences in test scores can be calculated using repeated measures ANOVA, paired sample t-test or the Wilcoxon signed rank test and the magnitude of the change can be reported using the effect size or a SRM. All of these are considered distribution-based methods of analysis, providing a way of expressing the observed change in a standardised metric (Hays et al. 2005). However, there is no indication about the level of importance of the changes from these parameters.

6.2.2.2 Anchor- or criterion-based methods of calculating responsiveness

The MCID is an example of a parameter of responsiveness that relies on clinically-observed changes and can be derived using the anchor or criterion-based method (Copay et al. 2007). A level of importance will be attributed to the observed change by the anchor or criterion. The criterion-based method uses a gold-standard comparator outcome measure that has already been found to be sensitive to changes in the population of interest. The anchor-based method collects information from either the patient or clinician, about whether they consider a change in the condition has occurred. Responses from either the criterion measure or the anchor change questionnaire, using a 5, 7 (or more) response Likert scale, are correlated with change scores collected from the outcome measure(s) under investigation. These change scores can be correlated with responses in the "better" and "worse" response categories in the anchor questionnaire, to establish what scores represent a "clinically" important change. Receiver operating characteristic (ROC) curves (Streiner et al. 2014) can be used with the outcome measure in question and these change questionnaires to establish their MCID values relative to sensitivity and specificity values (Turner et al. 2009), in the population being studied.

Quantifying the responsiveness of an outcome measure by establishing MCID values will assist in defining the ability of that outcome measure to detect a clinical meaningful change. Anchor- and criterion-based methods using ROC curve analysis are increasingly being used to establish the responsiveness of outcome measures in different study populations (Stratford et al. 1998, Bohannon and Glenney 2014, Castien et al. 2012) including pwLLAs (Rushton et al. 2015). As

there are no recognised gold-standard outcome measures in the measurement of physical function with pwLLAs, the anchor-based method was therefore chosen for this study.

As also previously outlined in Chapter 2, when assessing validity of an outcome measure, an existing measure that is considered the gold standard or criterion, is often used. If a gold standard measure is not available then a construct approach to validation may be adopted (also referred to as “hypothesis testing”). Similarly, the assessment of responsiveness may also rely on testing hypotheses in the absence of a gold standard. However, when assessing responsiveness, the hypotheses should be concerned with expected mean differences between changes in scores in groups, or with expected correlations between changes in scores on the outcome measure in question and the changes in scores on another instrument that is known to have adequate responsiveness in the same population (de Vet et al. 2011). Direction and magnitude of any observed changes or correlations are also stated in any hypotheses, in order that they can be proved or otherwise.

While it is recognised that distribution-based methods can demonstrate the ability of an outcome measure to measure change, these alone do not provide sufficient information for meaningful interpretation of the results regarding the responsiveness of a particular outcome measure (Beaton 2001). A statistically significant change may occur without the change being a clinically significant one (Husted et al. 2000). However, anchor- or criterion-based methods of calculating the MCID may be considered appropriate to assess responsiveness, as these methods compare the changes in the outcome measure under investigation with either: the change captured in the anchor questionnaire from the patient’s or clinician’s perspective on the clinical condition or; the change in the criterion measure that has been proved responsive in the population and clinical condition under investigation.

6.2.3 Current evidence for the responsiveness of physical function outcome measures used with pwLLAs

There is limited evidence of responsiveness for the top five outcome measures regularly used by AHPs across the UK during prosthetic rehabilitation. Results from the systematic review reported in Chapter 4 show that indices of responsiveness were reported for only 11 outcome measures used with pwLLAs, namely: 2MWT, Houghton scale, Rivermead Mobility Index, LCI / LCI-5, ICF checklist, FIM, OPCS,

Amputee Activity Score, Goal Attainment Scaling, Barthel and SIGAM. However, the strength of the evidence presented, in all studies reporting responsiveness, was considered "unknown", because the statistical analyses used in the studies was considered inappropriate according to published criteria (Terwee et al. 2007).
The same six outcome measures; the SIGAM, LCI-5, TUG, 2MWT, SCS, EQ-5D-5D, that were investigated in the Reliability Study reported in Chapter 5 were also investigated in this study. The published evidence on the responsiveness of these outcome measures used with pwLLAs is limited, see Table 6.1.

Rushton et al (2014) reported the MCID of the L-test, a modified TUG, for individuals with lower limb amputation in a recent study (n=33). However, clinically important differences have not been established for any of the outcome measures used regularly by AHPs involved in prosthetic rehabilitation in the UK (chapter 2). Therefore, the aim of this study was to establish the responsiveness of physical function outcome measures regularly used with lower limb amputees when assessing change during the rehabilitation period immediately following limb fitting.

Table 6.1 Responsiveness data for outcome measures used in this current study

Outcome Measure	Study	Responsiveness parameters presented	Level of Evidence / Methodological Quality
SIGAM	Ryall et al (2003a)	Effect size = 10.66	? (choice of analysis) / Fair (details of how missing data was handled not given, Apriori hypotheses not fully described, effect size only)
LCI / LCI-5 / PPA (LCI section)	Franchignoni et al (2004)	Wilcoxon signed rank showed significant difference for both LCI and LCI-5 Effect size: LCI = 1.09, LCI-5 = 1.40	? (choice of analysis) / Fair (effect size only)
	Rushton et al (2002)	Effect size LCI = 3.7	? (choice of analysis) / Fair (details of how missing data was handled not given, interim period not described & effect size only)
TUG	None reported in this population		
2MWT	Brooks et al (2001)	ANOVA ($p<0.001$) calculated for distances walked before and after rehabilitation	? (choice of analysis) / Fair (details of how missing data was handled not given, interim period not described & ANOVA only)
SCS	Hanspal et al (2003)	Wicoxon z analysis showed sensitivity to changes: 1. after socket adjustment (n=22) $z = -4.16$ $p<0.001$ 2. after delivery of new socket (n=5) $z = -2.06$ $p<0.05$	? (choice of analysis) / Fair (effect size only)
EQ-5D-5L	None reported in this population		

6.3 Methods

6.3.1 Participants

Adult (18 years or older) unilateral lower limb amputees, at either TT or TF level, who were receiving their first prosthetic limb were eligible for inclusion in the study.

Amputees were excluded if they were going to be receiving a prosthesis for transfer activities only. Patients with co-morbidities that prevented them undertaking any of the physical activities involved in the outcome measures, or if they had poor cognition or an insufficient comprehension of the English language that prevented them from understanding the written questionnaires, were also excluded from participating in the study. These exclusions were discussed with participants and their clinical team before the researcher made the final decision on who was included.

6.3.2 Ethical opinion

Favourable ethical opinions were also gained from both Queen Margaret University Ethics Committee and the NHS Ethics Committee (South East Scotland 02 committee) before commencing this study. In addition, NHS Lothian management approval was sought and gained prior to conducting the study on NHS Lothian premises.

6.3.3 Recruitment process

Amputees receiving prosthetic rehabilitation with their primary prosthesis at Astley Ainslie Hospital in Edinburgh, and fitting the inclusion criteria, were prospectively recruited into the study.

After the decision to fit a prosthesis was made by the multi-disciplinary team (MDT), comprising; physiotherapists, occupational therapists, prosthetists, nurses and a Consultant in Rehabilitation Medicine, potential participants were identified by the physiotherapist, according to the above inclusion and exclusion criteria.

Any amputees who matched the study criteria were given a Study Information Pack (SIP). The SIP included an invitation letter, a Participant Information Sheet (PIS) and a copy of the consent form they would be asked to sign.

The researcher was available to visit the amputees on the hospital ward, to answer any questions and discuss the study further with them before they decided to participate. If the amputee decided to take part, the physiotherapist arranged a suitable date and time with the researcher for the first study visit (SV1) when the amputee was enrolled as a participant into the study. Written consent was obtained from the participant at this visit which was always at least 24 hours after the amputee first received the SIP.

The flow chart (Figure 6.1) outlines the decision process that was followed during recruitment.

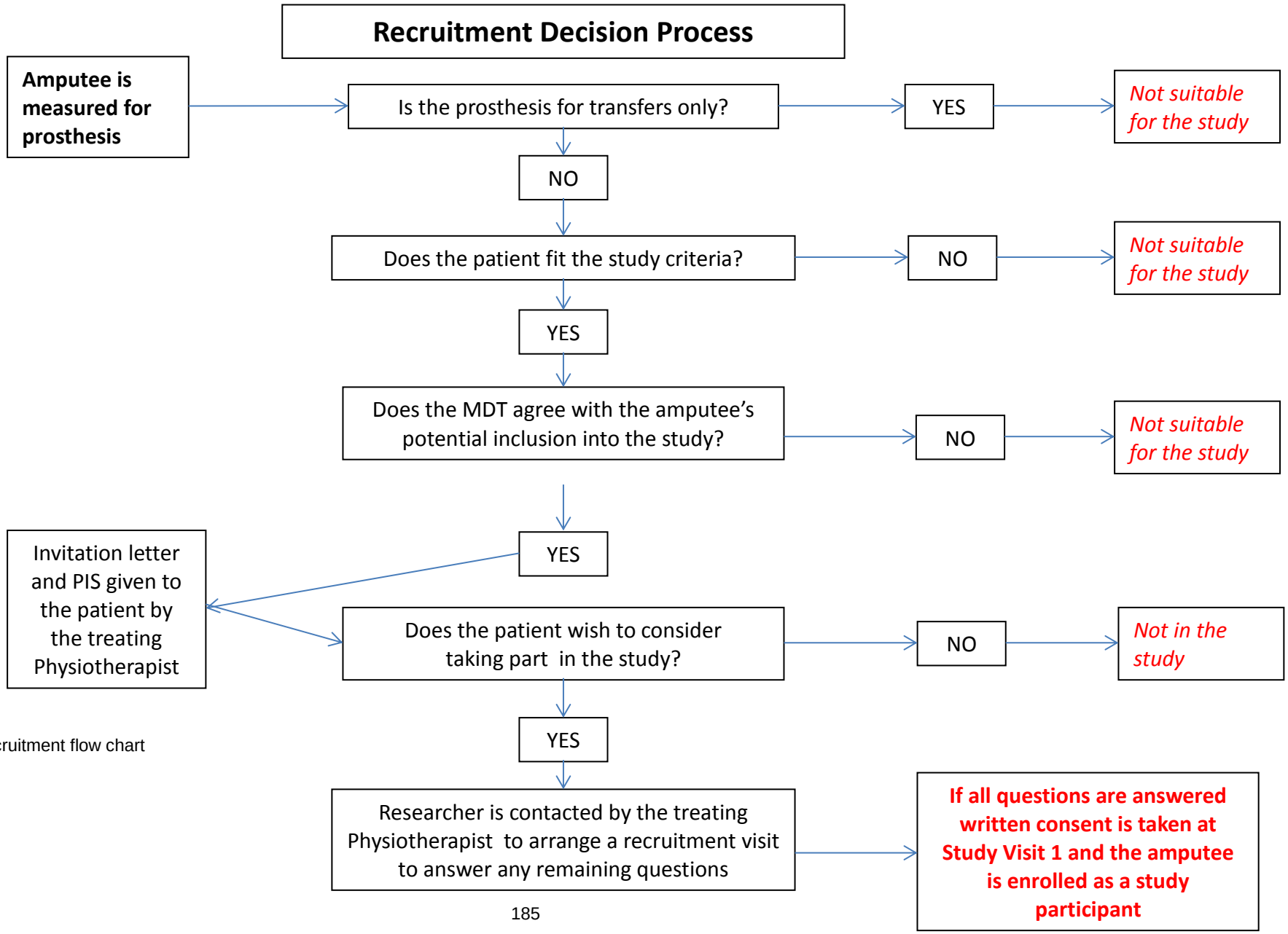


Figure 6-1 Recruitment flow chart

6.3.4 Study protocol

All study visits took place in the main gym and long corridor (walking tests) and a side room (PROMs) in the Physiotherapy Department at Astley Ainslie Hospital, Edinburgh. The study visits were planned to align with the model-of-care delivered for pwLLAs at this rehabilitation facility. See figure 6.2 for the individual participant pathway and Table 6.2 for a breakdown of activities at each study visit.

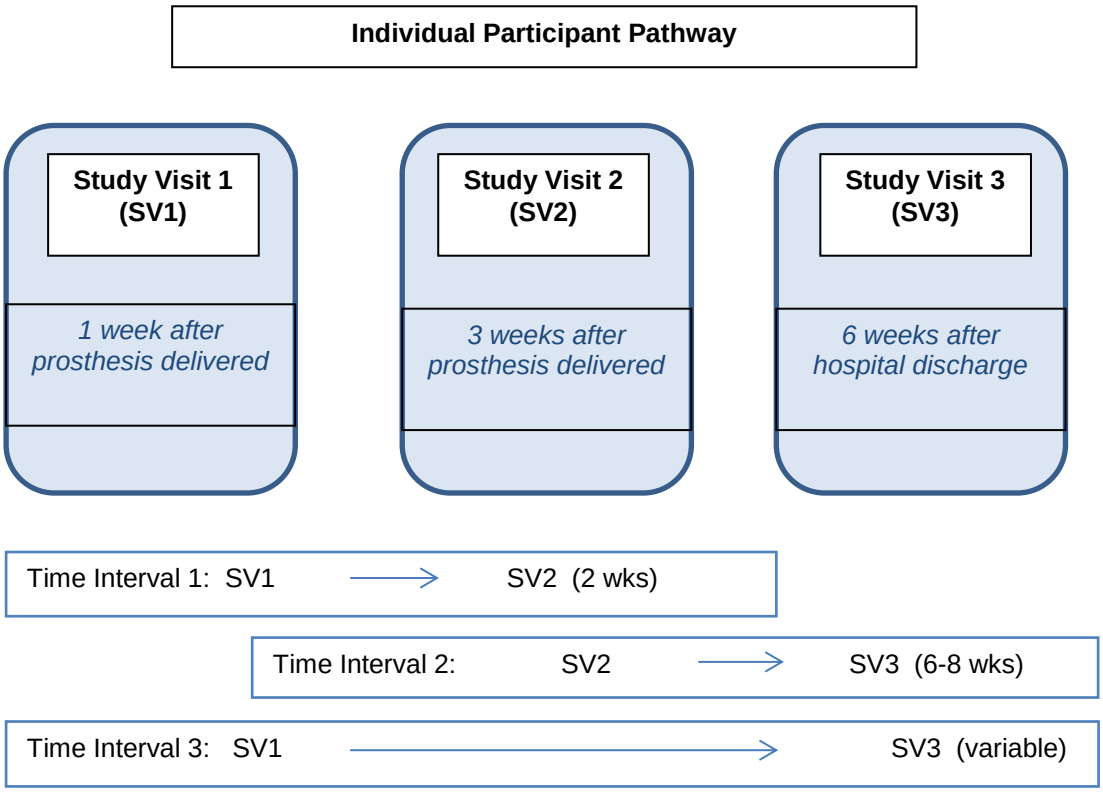


Figure 6-2 Individual participant pathway

Time Interval 1 (TI1) represents the period after delivery of the primary prosthetic limb and initial gait training and therapy. Time Interval 2 (TI2) represents the immediate post-hospital discharge period and Time Interval 3 (TI3) represents the early prosthetic rehabilitation for a pwLLA which was the focus for this study.

Table 6.2 Study visit details

Activity	Details	When / Where	Who
Identification of Study Participants	Identification of potential study participants Provision of Information Sheet	Prosthetic decision taken at hospital or in clinic	Health Professional (Physiotherapist)
Recruitment Visit (*optional*)	Explanation of Study Protocol and answer any questions	During the week before delivery of prosthesis at hospital	Researcher
Study Visit 1 (SV1)	Enrolment of participant - written consent obtained Basic Demographic details Baseline measurements for all outcome measures	1 week after prosthesis fitted at hospital	Researcher
Study Visit 2 (SV2)	Repeat outcome measures and Activity Change Questionnaire (ACQ)	3 weeks after prosthesis fitted at hospital	Researcher
Interim Study Visit (SVi)	Repeat outcome measures (*not done if discharge occurred within 1 week of SV2*.)	3-5 days prior to discharge from hospital	Researcher
Study Visit 3 (SV3)	Repeat outcome measures and ACQ	6 weeks post-discharge at clinic visit	Researcher

At the first Study Visit (SV1) written informed consent was obtained from the amputee by the researcher and the amputee was enrolled as a study participant. Baseline measurements for all the outcome measures (see below) were taken at this visit as well as basic demographic data recording: sex; age; date, cause and level of amputation; date of prosthetic fitting and, any relevant concomitant medical history.

Study Visit 2 (SV2) occurred two weeks following SV1 and the outcome measures were repeated. Every effort was made to complete the outcome measures at the same time of day for each of the participant's study visits and the order of the measurements was randomly assigned at each visit. An Activity Change Questionnaire (ACQ) was given to the participant to record how they felt their physical ability to perform everyday tasks had changed compared to when they first got their prosthetic limb (see later section on outcome measures for details).

The final visit, Study Visit 3 (SV3) coincided with the participant's six-week post-discharge MDT clinic appointment at the Southeast Mobility and Rehabilitation

Technology (SMART) Centre. The SMART centre is adjacent to the Physiotherapy Department at Astley Ainslie Hospital and the study data was collected in the same location as the previous visits. The outcome measures were repeated at this visit and in addition two further ACQs were given to record how the participants felt their physical ability to perform everyday tasks had changed compared to: i) when they were discharged from hospital and; ii) when they first got their artificial limb.

6.3.5 Outcome measures

The same outcome measures that were used in study III (the repeatability study reported in Chapter 5) were used in this study: SIGAM, LCI-5, TUG, 2MWT, EQ-5D-5L™. Full details of the outcome measures were given in Chapter 5 (section 5.3.5). In addition, an Activity Change Questionnaire (ACQ) was given to the participants to record how they felt they had changed, with respect to their physical function, between study visits. Data collection sheets and a detailed explanation of the procedures followed for each outcome measure in this study are included in Appendix 10.

6.3.5.1 Activity Change Questionnaire (ACQ)

The change questionnaire used in this study, was devised by the Researcher based on the five level response scales used in the SF-36 questionnaire asking patients to rate their general health (Ware and Gandek 1998). However, the leading question associated with the change questionnaire used in this study, was related to their physical abilities. Activity Change Questionnaires were presented to participants at both SV2 and SV4. By circling one of the statements: much worse, worse, the same, better, or much better; participants were asked to indicate “how you feel your current physical ability to perform everyday tasks is today compared to …..”. At SV2 participants were asked to compare their physical ability to when they first received their artificial leg. At SV4 they were asked the same question. In addition there were asked how they felt their current physical ability to perform everyday tasks compared to when they left hospital (SV2).

6.3.6 Statistical analysis

The method section in Chapter 5 (section 5.3.6) described the choice for considering the data on the TUG, 2MWT, LCI-5 EQ-5D-5L index score, EQ-5D-VAS and SIGAM as continuous data, while the SCS data were considered as ordinal.

6.3.6.1 Statistical power

An a priori sample size calculated showed that a sample size of n=30 would be sufficient to detect a statistically significant difference (2 sided α=0.05, power = 80%) between the outcome measures at SV1 and SV2 with an effect size of 0.56 or higher (Hedeker et al 1999).

6.3.6.2 Missing data and imputation analysis

The total time, distance or scores recorded for each outcome measure was considered a data point and an individual question in any questionnaire was considered a data item. All collected data was inspected at the time of collection. The data was inspected for missing data points and percentages were reported using SPSS. Multiple imputation methodology is recommended for longitudinal studies (Schafer and Graham 2002), to maintain sample size in the statistical analysis and to minimise the risk of bias. There were no missing data items in any of the questionnaire data collected and therefore multiple imputation was not required at the item level (Eekhout et al. 2014). There were missing data points at SV2 (3/30) and at SV3 (6/30), though not the same participants. SPSS was unable to run simulations on the missing data relative to the data available using the multiple imputation (regression method) analysis because there were too few unique data points available, i.e. some patients achieved the same scores. Therefore, no multiple imputation was performed and descriptive and inferential statistical analyses were performed on the original data set.

6.3.6.3 Summary descriptive analysis

The Shapiro Wilks test was used to determine whether the data were normally distributed and thus informing the type of inferential statistics to be performed.

Summary descriptive statistics were derived: means and standard deviations (SD) for parametric data; and medians and inter-quartiles for non-parametric and ordinal data (i.e. SCS).

6.3.6.4 Inferential Analysis

6.3.6.4.1 Significant differences across each time interval

Repeated measures ANOVA / Friedmans ANOVA were used to determine whether scores were significantly different over the three assessments for each outcome measure. Where significant differences were detected, post-hoc tests (Paired t-test with Bonferoni adjustment / Wilcoxon signed rank test) were carried out to determine between which assessments statistically significant differences existed. A level of $p<0.05$ (adjusting for Bonferoni $p=0.05/3 = 0.017$) was considered statistically significant

6.3.6.4.2 Effect size calculations across each time interval

The magnitude of change for each outcome measure was calculated using Cohen's d effect size across each time interval using the following formula:

$$d = \frac{M_1 - M_2}{SD\ pooled}$$

where M_1 = group 1 mean, M_2 = group 2 mean, *SD* = standard deviation, and

$SD\ pooled = \sqrt{[(SD_1^2 + SD_2^2)\ /2]}$

If $d < 0.5$ then the effect size is considered to be "small", $0.5 \leq d < 0.8$ represents a "medium" effect and $d \geq 0.8$ represents a "large" effect size (Cohen 1988).

6.3.6.4.3 Amputee perception of change

The anchor-based method of calculating the responsiveness of each of the outcome measures utilised the ACQ as the anchor. The ACQ collected the subjective perception of change by the participants. These scores were then compared to the change scores of the outcome measures tested, for each time interval i.e. T1, T2 and T3. How often each outcome measure correctly identified the participants who had and had not undergone an important change, according to the ACQ, was also quantified in each of the time intervals.

6.3.6.4.4 Validity of the ACQ

In addition to the statistical analyses performed on each of the outcome measures, the following criteria were used to assess the validity of the ACQ. The criteria used to established validity were based on those used in previous studies (Turner et al. 2009, Rushton et al. 2015). The ACQ was assessed to ensure that it is a valid tool to assess change in the constructs measured by each of the outcome measures.

For the ACQ to be deemed an adequate tool the following criteria must be true:

- the correlation of the ACQ with the difference between follow-up (SV2 and SV3) and baseline (SV1) outcome measure scores is more than 0.5 in absolute value, and;
- there is a negative correlation between the ACQ and the baseline outcome measure scores (SV1) in each time period, and;
- there is a positive correlation between the ACQ and the follow-up outcome measure scores (SV2 and SV3) in each time period, and;
- the correlation of the ACQ with the difference between follow-up (SV2 and SV3) and baseline (SV1) outcome measure scores, is at least 0.2 greater (in absolute terms) than the correlations with either the baseline or the follow-up score.

6.3.6.4.5 Receiver Operating Characteristic (ROC) curve analysis

To determine the accuracy of each outcome measure in detecting a clinically meaningful change (as defined by the ACQ) a ROC curve analysis was performed for each outcome measure at each time interval. This analysis requires a dichotomous split in the results where a cut off point is identified; in this case the point where a clinically meaningful change had occurred with reference to the ACQ scores. In this instance, the two groups the participants were split into in each time interval were: those who recorded a rating of 5 on the ACQ and all the remaining participants. A rating of 5 (“much better”) on the ACQ was used to identify participants who were considered to have undergone an important (clinically meaningful) change during each study time interval. The remainder, those who reported “better” (4), “same” (3), “worse” (2), or “much worse” (1) on the ACQ, were considered not to have made an important change. Therefore a cut-off point between those who were in a “positive actual state” or the “much better” group and those in a “negative actual state” or the “others” group was identified.

A ROC curve is produced by plotting sensitivity on the y-axis against 1-specificity on the x-axis (Streiner et al. 2014). In the context of this study, sensitivity refers to the number of participants correctly identified by the outcome measure as having undergone an important change, divided by all participants who perceived that they underwent an important change, according to the ACQ. Specificity refers to the number of participants who were correctly identified by the outcome measure as not undergoing an important change divided by all participants who perceived that they did not undergo an important change, according to the ACQ. To quantify how accurate each outcome measure was in correctly identifying individuals who had and had not undergone an important change, an area under the curve (AUC) was calculated.

See Figure 6.3 for an example graph demonstrating high (AUC> 0.90), moderate (0.70 – 0.90) accuracy and no better than chance result (<0.50) (Fischer et al. 2003).

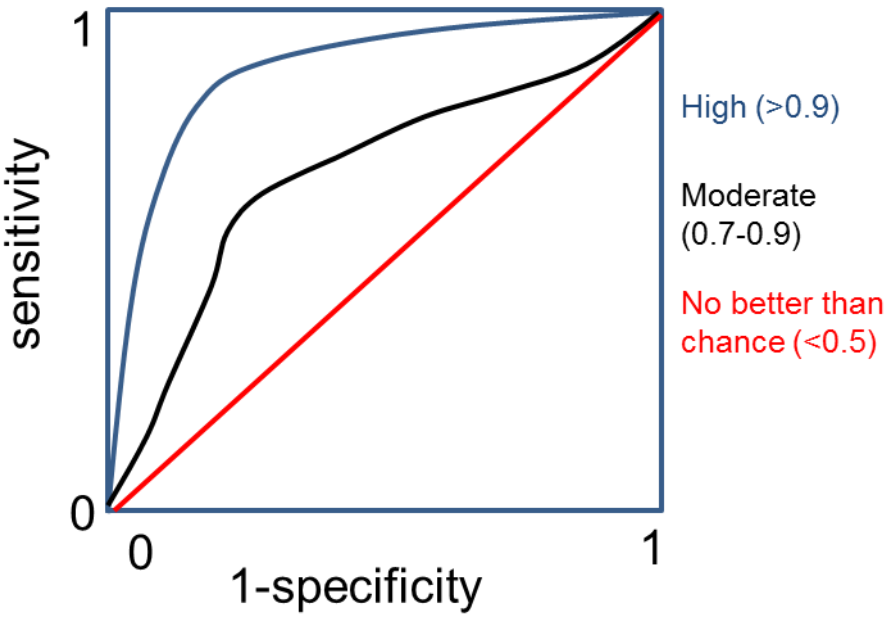


Figure 6-3 Example ROC curve graph

6.3.6.4.6 Minimal Clinically Important Difference (MCID)

The MCID was defined as the change score that best distinguished between those participants who had and had not reported that they were much improved on the ACQ. The MCID was identified using the coordinate closest to the upper left hand corner of the ROC curve and its corresponding sensitivity and specificity values (Turner et al. 2009).

6.3.6.4.7 Construct approach to measuring responsiveness

In addition to the analysis that described both clinically and statistically significant change scores and the magnitude of those changes, a construct approach ("hypothesis testing") to measuring responsiveness was also taken (de Vet book). Specific a priori hypotheses were therefore formulated.

It was hypothesised that;

1) participants who rated themselves as having undergone a greater improvement in their physical function (i,e, higher ACQ scores) will show a superior improvement in the six outcome measures over the same time interval; i.e. they will have greater change scores in the SIGAM, LCI-5, SCS and EQ-5D-index and EQ-5D-VAS, greater change times in the TUG and greater change distances in the 2MWT.
2) the magnitude of the change in the six outcome measures would be larger in those who perceived that their physical function was "much better" (ACQ score = 5) compared to who did not. (ACQ scores 1-4) and:
3) the outcome measures would correctly identify individuals who had and who had not undergone an important change (i.e. ACQ of 5 vs ACQ 1-4) 80% of the time.

To address hypothesis 1; it was expected that there would be a correlation of at least 0.6 between the change in the instrument scores and the patient perceived change in physical function, as measured using the ACQ, for each time interval (1, 2 and 3). The expected sign of the correlation will be negative for the TUG, and positive for all other outcome measures, as a decrease in TUG times denotes an improvement.

To address hypothesis 2; it was expected that there would be a difference in the effect size of at least one "level" (i.e. low to medium, or medium to high, as denoted by the criteria presented by Cohen (1988) during each time interval (1, 2 and 3) for those amputees who rated 5 on the ACQ vs those who rated 1-4. It was also expected that the group mean of the "improved greatly" group will be larger than the group mean of the rest, for each outcome measure, in each time interval.

To address hypothesis 3; it was expected that the AUC on a ROC curve analysis performed for each outcome measure for each time interval would achieve at least moderate accuracy i.e. 0.8, as interpreted using the following criteria: AUC of >0.9 = high accuracy; 0.7–0.9 = moderate accuracy; 0.5–0.7 = low accuracy; and ≤ 0.5 = chance result (Fischer et al. 2003).

6.3.6.5 Statistical Software

All descriptive and inferential statistics were calculated using IBM SPSS version 21.0.0.1 (2012).

6.4 Results

6.4.1 Participants

6.4.1.1 Demography and aetiology of study population

A total of 124 amputees were screened during a twenty-two month recruitment period, as they passed through the rehabilitation hospital following their amputation surgery. Of these, 83 did not meet the inclusion criteria. Table 6.3 details the reasons why these patients were not suitable for inclusion. A further six patients did not give consent and for five the researcher was not available, hence a final total of thirty patients were recruited into the study. Table 6.4 presents the demographic and basic aetiological data for the thirty participants.

Table 6.3 Reasons for non-inclusion

Reasons	**Number of amputees**
Not limb-fitted	29
Bilateral	18
Unfit physically	15
Not a primary amputation	8
Discharged less than a week following fitting	7
Unfit cognitively	6
Total not suitable for inclusion	83

Table 6.4 Demography and basic aetiology

Age - yrs	mean (SD)	61 (13.7)	
	median	61	
Gender		n	%
	Male	24	80
	Female	6	20
Level	Trans-tibial (TTA)	22	73
	Trans-femoral (TFA)	8	27
Side	Right	13	43
	Left	17	57
Main Cause of amputation	PAD	13	43
	Diabetes	8	27
	Trauma	2	7
	Other	7	23
Number of co-morbidities	≤ Three	27	90
	≥ Four	3	10

The mean age of participants in this study (61 years) was similar to the population recruited in the Repeatability study reported in Chapter 5. This is again slightly younger than the mean age of 67 years reported for the amputee population by the SPARG annual report for 2013 (Scott et al. 2016). As previously observed in chapter 5, only 40% of those patients undergoing a lower limb amputation are limb-fitted and SPARG does not report on the ages of those amputees who are limb-fitted separately, but who are likely to be younger. The male to female ratio was slightly higher (4:1) in this study than that reported in the SPARG report, which was 67% male to 33% female. The proportion of TT amputees to TF amputees in the current study (73% vs. 26%) shows a slightly higher percentage of TT amputees to the percentages reported by SPARG. The 2013 report stated that 65% of all limb-fitted amputees were at the TT level, 23% were at the TF level with the remainder at other levels.

The main cause of amputation in this study was either peripheral arterial disease or diabetes, which presents a typical profile of the population of amputees across Scotland. However, the total percentages reported for these causes in the SPARG report is slightly higher with a combined total of 85% of all amputations compared to

70% in the current study. As stated previously in Chapter 5, it is conceivable that pwLLAs who had an amputation because of trauma or other causes may be higher in the limb-fitted group. The aetiology breakdown of participants in this study is more similar to the pattern of lower limb referrals to prosthetic centres across the UK, as seen in the limbless statistics report for 2011-12 (UNIPOD 2013). This report showed that referrals for all dysvascular reasons and diabetes accounted for 57% of the total, and referrals for trauma and other reasons accounted for 10% and 33%, respectively.

6.4.1.2 Walking aids used

Of the 27 participants on whom data was collected at SV2, 11 demonstrated they were less dependent on walking aids compared with SV1, e.g. they were using one stick or crutch rather than two, see Figure 6.4. Twelve participants recorded a decrease in walking aid use between SV2 and SV3 and 18 across the whole study period, thus demonstrating that participants were improving between each visit.

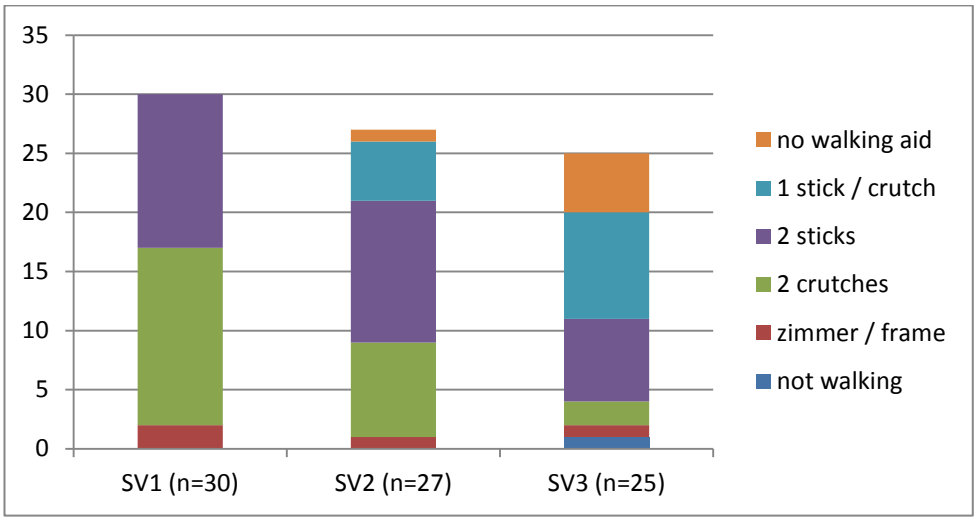


Figure 6-4 Walking aids used at each study visit

6.4.2 Study time periods

The time periods to key milestones and for each time interval are displayed in Table 6.5. The range of days from amputation to fitting included four participants for whom the period from amputation to fitting was greater than 100 days. However, there did

not appear to be any relationship between this and the results from the outcome measures.

Table 6.5 Time periods in days

	Amputation to fitting	Fitting to SV1	Fitting to discharge	**SV1 to SV2 Time Interval 1**	**SV2 to SV3 Time Interval 2**	**SV1 to SV3 Time Interval 3**
mean	70	9	24	**12**	**74**	**84**
SD	65	3	19	**4**	**37**	**37**
range	25-351	5-16	2-111	**6-21**	**42-184**	**49-195**
median	47	8	22	**12**	**63**	**74**

6.4.3 Data collection issues

There were some issues at both SV2 and SV3, which affected data collection.
One participant had a delayed hospital discharge due to housing problems and therefore all his / her study visits were undertaken while this participant remained in hospital. Study visit 3 was undertaken approximately 6 weeks after SV2.

Due to ongoing medical issues, one participant was readmitted after SV2, however, the participant was seen by the multi-disciplinary team for their 6 week "post-discharge" review appointment while in hospital at approximately 6 weeks after their SV2. Study data was collected at this time point for SV3. Full details of all the issues and their impact can be found in Tables 6.6. and 6.7

Table 6.6 Issues at SV2

Participant	Issue	Impact on SV2
P02	**Patient feeling unwell**	**SV2 not completed**
P03, P06 P10	Discharged home prior to SV2, planned attendance as an out-patient	No impact – SV2 data collected approx. 2 weeks after SV1
P13	**Discharged home prior to SV2, no capacity for attendance as an out-patient**	**SV2 not completed**
P21	**Discharged home prior to SV2, planned attendance as an out-patient, injury to residuum and patient feeling unwell.**	**SV2 not completed**

Table 6.7 Issues at SV3

Participant	Issue	Impact on SV3
P02	**Did not attend (DNA) 1st clinic visit and not wearing limb at 2nd clinic visit due to fit issues**	**Unable to collect SCS, TUG or 2MWT data**
P07 **P30**	**Readmitted for further amputation surgery. Now bilateral, withdrawn from study.**	**SV3 not completed**
P09	Readmitted to hospital at time of post-discharge clinic appointment	No impact - SV3 data collected while in hospital
P16	**1st visit DNA - rescheduled but unable to attend as in acute hospital with further complications, therefore withdrawn from study**	**SV3 not completed**
P18	**Transport turned up and patient left clinic before completing the walking tests**	**Unable to collect TUG or 2MWT data**
P26	Remained an in-patient at time of the 6 week post therapy discharge	No impact - SV3 data collected approx. 6 weeks after SV2
P27	**Researcher unable to attend clinic.**	**SV3 not completed**

6.4.4 Missing data

There were no missing data items in any of the questionnaire data collected. However, due to the issues at SV2 and SV3 described above, some data points were missing at both visits. Percentages of missing data points for each outcome measure are listed in Table 6.8.

Table 6.8 Results from missing data point analysis by outcome measure

Variable	Missing / Total (% missing)		
Demographic data	None missing		
Outcome Measures	SV1	SV2	SV3
SIGAM	None missing	3 / 30 (10)	4 / 30 (13)
TUG		3 / 30 (10)	6 / 30 (20)
2MWT		3 / 30 (10)	6 / 30 (20)
LCI-5		3 / 30 (10)	4 / 30 (13)
SCS		3 / 30 (10)	5 / 30 (17)
EQ-5D index score		3 / 30 (10)	4 / 30 (13)
EQ-5D VAS		3 / 30 (10)	4 / 30 (13)
ACQ		3 / 30 (10)	8 / 60 (13)

6.4.5 Summary descriptive analysis results

Data collected for SIGAM, LCI-5, TUG, 2MWT, and EQ-5D-5L was assessed using Shapiro Wilks in order to determine whether the data were normally distributed. It was found that only the EQ-5D-index data was normally distributed at all three time points (see Appendix 13). Summary statistics were derived, as appropriate, for all available data collected at each time point and are presented in the following table (Table 6.9).

Table 6.9 Summary descriptive statistics all study visits

Outcome measure (units)	**Summary Statistics**	**SV1 n=30**	**SV2 n=27**	**SV3 n= 24†, 25††, 26**
SIGAM (grades)	Median [IRQ]	4.0 [4.0 – 4.0]	8.0 [4.0 – 8.0]	8.0 [4.8 – 9.0]
	Range	2-8	3-9	1-10
LCI-5 (points)	Median [IRQ]	35.5 [25.0-41.25]	38.5 [35.0-44.0]	43.5 [38.7-47.3]
	Range	6-45	19-52	18-54
TUG (secs)	Median [IRQ]	30.7 [22.4-46.9]	21.8 [16.3-41.9]	15.7 [10.4-26.0]
	Range	15.1-84.3	9.2-62.6	7.9-41.1
2MWT (m)	Median [IRQ]	52.1 [35.0-62.4]	68.8 [42.7-73.3]	83.7 [50.2-112.7]
	Range	20.9-86.2	25.6-137.3	32.8-172.9
SCS (points)	Median [IRQ]	7 [6-8]	8 [6-8]	8 [7-9]
	Range	3-10	3-10	3-10
EQ-5D- index (points)	Mean (SD)	0.654 (0.186)	0.755 (0.132)	0.679 (0.205)
	Range	0.221-1.000	0.401-1.000	0.057-1.000
EQ-5D- VAS (points)	Median [IRQ]	80 [70-90]	90 [90-90]	80 [80-90]
	Range	50-100	45-100	50-97

† sample size for TUG & 2MWT at SV3, †† sample size for SCS at SV3

The group mean differences recorded across each time interval are presented in Table 6.10. The different sample size at each study visit is reflected in different paired sample sizes in TI2 and TI3 for some outcome measures.

Table 6.10 Group mean differences (SD) across each time interval

Outcome Measure (units)	Time Interval 1 SV1-SV2 n=27	Time Interval 2 SV2-SV3 n=22†, 23	Time Interval 3 SV1-SV3 n=24‡, 25‡‡, 26
SIGAM (grades)	2.0 (2.22)	0.8 (2.5)	2.5 (2.3)
LCI-5 (points)	6.3 (7.3)	4.1 (6.9)	10.9 (8.7)
TUG†‡ (secs)	-8.3 (7.2)	-6.3 (7.1)	-12.9 (9.3)
2MWT†‡ (m)	16.3 (20.8)	15.7 (22.3)	34.7 (34.4)
SCS†‡‡ (points)	0.22 (2.1)	0.61 (2.2)	0.81 (2.1)
EQ-5D index (points)	0.091 (0.163)	-0.026 (0.167)	0.052 (0.259)
EQ-5D VAS (points)	3.6 (10.8)	-3.7 (12.0)	0.7 (9.5)
† sample size for TUG, 2MWT & SCS in TI2, ‡ sample size for TUG & 2MWT at TI3, ‡‡ sample size for SCS at TI3			

6.4.6 Inferential analysis

6.4.6.1 Significant differences across each time interval

The group mean differences indicated improvements for all outcome measures over all time intervals except for the EQ-5D-index and EQ-5D–VAS scores during TI2 (post-hospital discharge).

There were statistically significant differences recorded for all outcome measures except the SCS, EQ-5D-index and EQ-5D-VAS, as determined by repeated measures ANOVA for parametric data and Friedman's ANOVA for non-parametric data. These results and those of the post-hoc tests are presented in Table 6.11. Post-hoc tests were not done for SCS or EQ-5D-index scores, but for those done they revealed that there were statistically significant differences between all the study visits for LCI-5, TUG and 2MWT.

6.4.6.2 Effect size calculations across each time interval

The magnitude of change for each outcome measure was estimated by calculating Cohen's d effect size, using pooled SD, across each time interval. If d < 0.5 this is considered a "small" effect, 0.5 ≥ d < 0.8 a "medium" effect and d ≥0.8 represents a "large" effect size (Cohen 1988). Therefore, in the table below (Table 6.11) effect sizes <0.5 have been labelled as "small', ≥ 0.5 and <0.8 as "medium"and ≥0.8 as "large".

Table 6.11 Repeated measures ANOVA / Friedman's with post-hoc analysis and effect size per outcome measure for each time interval

Outcome measures	Friedman's ANOVA except EQ-5D-index: Repeated Measures ANOVA	Post-hoc tests Wilcoxon Signed Ranks			Cohen's d		
		Time Interval 1 SV1-SV2 n=22†, 23	Time Interval 2 SV2-SV3 n=22†, 23	Time Interval 3 SV1-SV3 n=22†, 23	Time Interval 1 SV1-SV2 n=27	Time Interval 2 SV2-SV3 n=22†, 23	Time Interval 3 SV1-SV3 n=24‡, 25‡‡, 26
SIGAM	$X^2(2)=24.356$, p<0.001	<0.017*	0.156	<0.017*	1.03	0.35	1.13
LCI-5	$X^2(2)=30.615$, p<0.001	<0.017*	<0.017*	<0.017*	0.66	0.53	1.13
TUG	X^2 (2)=31.182, p<0.001	<0.017*	< 0.017*	<0.017*	-0.53	-0.54	-1.12
2MWT	X^2 (2)=21.402, p<0.001	<0.017*	<0.017*	<0.017*	0.69	0.47	1.20
SCS	X^2 (2)=4.831, p=0.089	Not required p>0.05			0.13	0.36	0.49
EQ-5D-index	F(2,44)=2.672, p= 0.08	Not required p>0.05			0.61	-0.20	0.25
EQ-5D-VAS	X^2 (2)=12.585, p=0.002	0.019	0.087	0.685	0.31	-0.34	0.06
		† sample size for TUG & 2MWT *=statistically significant at p<0.017 in post-hoc tests, adjustments for multiple comparisons: Bonferoni			† sample size for TUG, 2MWT & SCS in TI2, ‡ sample size for TUG & 2MWT at TI3, ‡‡ sample size for SCS at TI3 d<0.5 = "small", d≥0.5, <0.8 = "medium" and d ≥0.8= "large" effect size		

Activity Change Questionnaire

The individual participant's perception of change scores collected at SV2 and SV3 are shown in Figure 6.5.

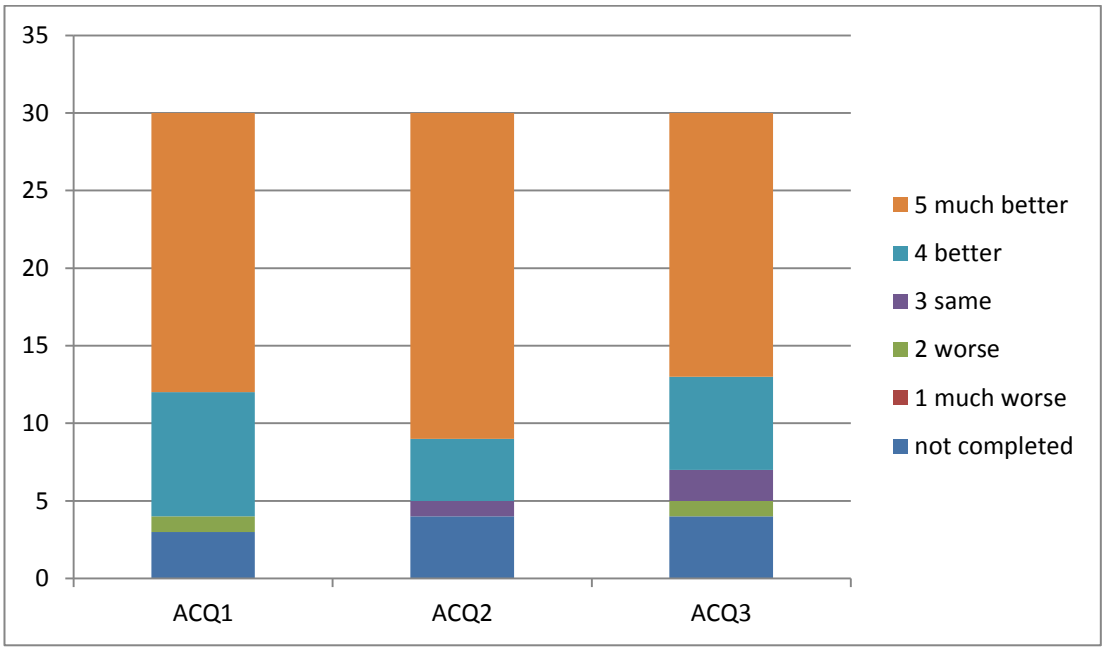


Figure 6-5 ACQ scores

6.4.7 Validity of the ACQ

The results for all the correlations of: i) the ACQ with the difference between follow-up and baseline outcome measure scores and; ii) between the ACQ and the baseline outcome measure scores, and; iii) between the ACQ and the follow-up outcome measure scores, and; between the ACQ and the difference between iv) follow-up and v) baseline outcome measure scores are presented in the following tables (Table 6.12 - TI1, Table 6.13 - TI2 and Table 6.14 - TI3).

All correlations are Kendall's tau except for the EQ-5D-index which are Pearson's product-moment, due to the parametric nature of the data.

The correlation criteria are listed in the first column and if the correlation is met for any of the outcome measures the result or a tick is highlighted in red in the table. For the ACQ to be considered a valid tool to measure perceived change then all criteria must be met. However, the ACQ did not satisfy all the criteria for any of the outcome measures across any of the time intervals.

Table 6.12 Time Interval 1 ACQ correlations

Correlation requirements	SIGAM	LCI5	TUG	2MWT	SCS	EQ5D index	EQ5D VAS
	n=27						
i) ACQvsdiff >0.5	.146	.324	-.213	-.053	-.253	.167	.013
ii) ACQvsSV1 -'ve correlation	**-.416**	**-.282**	.192	**-.148**	.475**	.283	.179
iii) ACQvsSV2 +'ve correlation	-.100	-.086	**.106**	-.156	**.100**	**.201**	**.295**
iv) ACQvsdiff – ACQvsSV1 >0.2	√	X	X	X	√	X	X
iv) ACQvsdiff – ACQvsSV2 >0.2	X	√	X	X	X	X	√
*p<0.05, **p<0.001 X = criteria not met, √ = criteria met							

Table 6.13 Time Interval 2 ACQ correlations

Correlation requirements	SIGAM	LCI5	TUG	2MWT	SCS	EQ5D index	EQ5D VAS
	n=23		n=22			n=23	
i) ACQvsdiff >0.5	.261	.166	.253	.288	-.089	-.269	.202
ii) ACQvsSV1 -'ve correlation	.195	.121	**-.270**	.367*	.143	.144	**-.069**
iii) ACQvsSV2 +'ve correlation	**.375**	**.264**	-.183	**.245**	-.083	-.153	**.121**
iv) ACQvsdiff – ACQvsSV1 >0.2	X	X	X	X	X	X	X
iv) ACQvsdiff – ACQvsSV2 >0.2	X	X	X	X	X	X	X
*p<0.05, **p<0.001 X = criteria not met, √ = criteria met							

Table 6.14 Time Interval 3 ACQ correlations

Correlation requirements	SIGAM	LCI5	TUG	2MWT	SCS	EQ5D index	EQ5D VAS
	n=26		n=24		n=25	n=26	
i) ACQvsdiff >0.5	.107	.141	-.005	-.089	-.035	.034	.159
ii) ACQvsSV1 -'ve correlation	**-.095**	**-.144**	.005	**-.016**	.156	**-.059**	.161
iii) ACQvsSV2 +'ve correlation	-.024	-.051	**0.47**	-.026	**.079**	-.010	**.350***
iv) ACQvsdiff – ACQvsSV1 >0.2	X	X	X	X	X	X	X
iv) ACQvsdiff – ACQvsSV2 >0.2	X	X	X	X	X	X	X
*p<0.05, **p<0.001 X = criteria not met, √ = criteria met							

6.4.8 Receiver Operating Characteristic (ROC) curve analysis

A ROC curve was produced by plotting sensitivity on the y-axis against 1-specificity on the x-axis. The sample sizes of the groups in each ROC analysis, taking into account any missing data, are detailed in Table 6.15. The same groups were utilised when testing all hypotheses.

Table 6.15 Group sample sizes of positive and negative actual states

<table>
<tr><td rowspan="2"></td><td colspan="3">Group sample sizes</td></tr>
<tr><td>Time Interval 1
SV1-SV2</td><td>Time Interval 2
SV2-SV3</td><td>Time Interval 3
SV1-SV3</td></tr>
<tr><td rowspan="2">Positive actual state =
ACQ response 5 =
“much better” group</td><td rowspan="2">All outcome measures
n=18</td><td>SIGAM, LCI-5,
EQ-5D-index & VAS
n=20</td><td rowspan="2">All outcome measures
n=17</td></tr>
<tr><td>TUG, 2MWT & SCS
n=19</td></tr>
<tr><td rowspan="3">Negative actual state =
ACQ response 1-4 =
“all others” group</td><td rowspan="3">All outcome measures
n=9</td><td rowspan="3">All outcome measures
n=3</td><td>SIGAM, LCI-5,
EQ-5D-index & VAS
n=9</td></tr>
<tr><td>SCS n=8</td></tr>
<tr><td>TUG & 2MWT
n=7</td></tr>
</table>

The ROC curve for LCI-5 at TI1 is presented in Figure 6.6. The AUC for this curve was calculated at 0.676 representing a low accuracy for the LCI-5 in this time interval. ROC curve graphs for all outcome measures across each time interval are presented in Appendix 14 and AUC results are presented in Table 6.16.

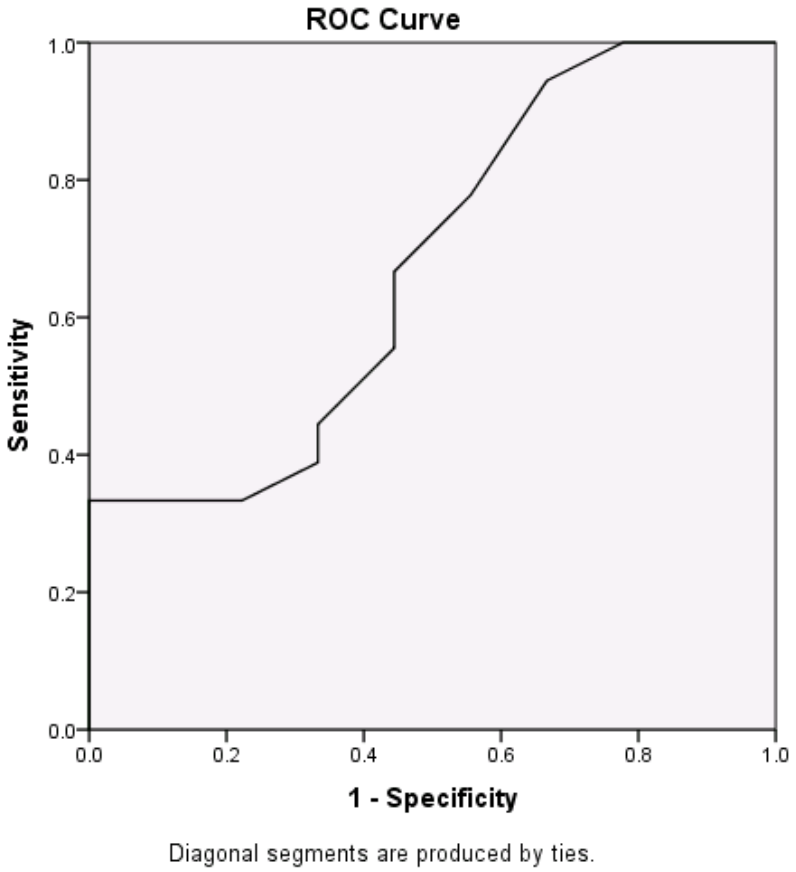


Figure 6-6 LCI-5 ROC curve for TI1 (from SPSS)

Table 6.16 ROC analysis area under the curve

Outcome Measure	Time Interval 1 SV1-SV2	Time Interval 2 SV2-SV3	Time Interval 3 SV1-SV3
SIGAM	0.565	0.742	0.546
LCI5	0.676	0.692	0.539
TUG	0.346	0.754	0.504
2MWT	0.466	0.789	0.445
SCS	0.330	0.417	0.507
EQ5D index	0.552	0.333	0.549
EQ5D VAS	0.485	0.700	0.572
Null hypothesis: true area = 0.5 or above AUC >0.9 = high accuracy; 0.7–0.9 = moderate accuracy; 0.5–0.7 = low accuracy; and ≤0.5 = chance result (Fischer et al 2003)			

It can be seen that both the SIGAM and LCI-5 have “better than chance” accuracy, over all time intervals, in measuring an important change, using the participant’s perception of change of physical ability as an anchor. The results obtained for the TUG and 2MWT indicated that the accuracy of these measures to detect an important change was at best, “moderate” in TI2 but was low or by chance in TI1 and TI3.

It should be noted there was at least one tie of scores occurring in both the positive actual state group and the negative actual state group for the SIGAM, LCI5, SCS, EQ-5D index and EQ-5D VAS in all time intervals, and for the 2MWT in TI1, which may bias the results.

6.4.9 Minimal Clinically Important Difference (MCID)

The MCID was identified using the coordinate closest to the upper left hand corner of the ROC curve and its corresponding sensitivity and specificity values, i.e. the coordinate where the highest sensitivity and lowest specificity scores met on the curve. (NB similar results regarding accuracy to those presented in Table 6.16)

Table 6.17 MCID values from the ROC analysis

Outcome Measure (units)	Time Interval 1 SV1-SV2	Time Interval 2 SV2-SV3	Time Interval 3 SV1-SV3
SIGAM (grades)	2.5	-1.0	4.5
LCI5 (points)	8.5	-1.5	21
TUG (secs)	-0.3	-12.4	-5.5
2MWT (m)	57.9	2.0	55.4
SCS (points)	2.0	1.5	2.5
EQ-5D index (points)	0.108	0.041	0.168
EQ-5D VAS (points)	12.5	-9.0	7.5
Null hypothesis: true area = 0.5 or above AUC >0.9 = high accuracy; 0.7–0.9 = moderate accuracy; 0.5–0.7 = low accuracy; and ≤0.5 = chance result (Fischer et al 2003)			

6.4.10 Construct approach (hypothesis testing) to assessing responsiveness

The group numbers in Table 6.15 were used throughout all the hypotheses testing analyses.

6.4.10.1 Hypothesis 1

"Participants who rated themselves as having undergone a superior improvement in their physical function (i,e, higher ACQ scores) will show a greater improvement in the six outcome measures over the same time interval; i.e. they will have greater change scores in the SIGAM, LCI-5, SCS and EQ-5D-index and EQ-5D-VAS, greater change times in the TUG and greater change distances in the 2MWT".

Pearson's correlation coefficients were calculated across the three time intervals, for the EQ-5D-index data which was normally distributed. Kendall's Tau correlation was used for the ordinal data obtained from the SCS and the non-parametric data obtained from the remaining outcome measures (see Table 6.18). The same groups, representing the positive and negative actual states, used in the ROC analysis (see Table 6.13) were used in the correlation calculations presented here.

Table 6.18 Correlations per outcome measure for each time interval

Outcome Measure	Time Interval 1 SV1-SV2		Time Interval 2 SV2-SV3		Time Interval 3 SV1-SV3	
	Pearson's	Kendall's Tau	Pearson's	Kendall's Tau	Pearson's	Kendall's Tau
SIGAM		.146		.261		.107
LCI-5		.268		.189		.061
TUG		-.213		.253		-.005
2MWT		-.053		.288		-.089
SCS		-.253		-.089		-.035
EQ-5D index	.167		.269		.034	
EQ-5D VAS		.013		.202		.159

No correlations that were statistically significant and above 0.6 were found and therefore the hypothesis was not confirmed for any of the six outcome measure. Graphs showing plots of the correlations for the outcome measures across all the time intervals are presented in Appendix 15

6.4.10.2 Hypothesis 2

"The magnitude of the change in the six outcome measures would be larger in those who perceived that their physical function was much better (ACQ= 5) compared to who did not (ACQ 1-4)."

In TI1 a greater effect size was noted for the SIGAM, LCI-5, TUG and EQ-5D index, see Table 6.19, for those amputees who rated themselves as having improved greatly ("much better" response on the ACQ), than for the others who felt they only slightly improved ("slightly better"), didn't improve ("the same") or were worse (either "slightly or much worse"). In TI2 only the LCI-5 and 2MWT demonstrated greater effect sizes in the group who perceived their physical function to be much improved, though there were only 3 participants in the "other" group. In TI3, only the SIGAM and EQ-5D index that recorded the greater effect sizes in the improved group.

Table 6.19 Effect size for each time interval per ACQ responses

Outcome Measure	Time Interval 1 SV1-SV2		Time Interval 2 SV2-SV3		Time Interval 3 SV1-SV3	
Group sample sizes	*Much better n=18*	*All others n=9*	*Much better n=20 n=19†*	*All others n=3*	*Much better n=17*	*All others n=9 n=8‡ n=7‡‡*
SIGAM	1.24	0.71	0.48	0.24	1.44	0.70
LCI-5	0.80	0.36	0.54	0.19	1.12	1.10
TUG†‡‡	0.58	0.40	0.44	1.32	1.05	1.17
2MWT†‡‡	0.64	0.77	0.48	0.09	1.18	1.15
SCS‡	0.22	0.70	0.30	0.51	0.44	0.57
EQ-5D index	0.58	0.47	0.32	0.41	0.62	0.14
EQ-5D VAS	0.28	0.43	0.25	1.39	0.12	0.05

d<0.5 = "small", d≥0.5 and <0.8 = "medium" d ≥0.8= "large" effect size

† sample size for "much better" group for TUG & 2MWT in TI2,

‡ sample size for "much better" group for SCS in TI3,

‡‡ sample size for "much better" group for TUG & 2MWT in TI3

Results for the other outcome measures show the opposite from what might have been expected: i.e during TI1 the 2MWT, SCS and EQ-5D VAS, during TI2 the TUG, SCS and EQ-5D index; and the SCS in TI3, the effect sizes were larger in the "other" group and not the group who perceived an improvement on the ACQ.

It was also expected that the group mean of the "much better" group will be larger than the group mean of the rest, for each outcome measure, in each time interval. As the difference in group means are highly related to the effect size (i.e. the same except they are divided by SD) they showed a similar pattern except for LCI-5 over TI3, where the group mean for the improved group was greater than the other group (see Table 6.20). The results are shown in **red** in the table where this was not the case.

Table 6.20 Group means for each time interval by ACQ responses

Outcome Measure (units)	Time Interval 1 SV1-SV2		Time Interval 2 SV2-SV3		Time Interval 3 SV1-SV3	
Group sample sizes	*Much better n=18*	*All others n=9*	*Much better n=20 n=19†*	*All others n=3*	*Much better n=17*	*All others n=9 n=8‡ n=7‡‡*
SIGAM (grades)	2.2	1.4	1.0	-0.7	2.8	1.9
LCI-5 (points)	7.8	3.1	4.4	1.0	11.4	10.0
TUG† ‡‡ (secs)	-9.8	-5.4	***-5.1***	***-12.0***	***-12.0***	***-15.0***
2MWT† ‡‡ (m)	***13.7***	***21.5***	17.1	1.4	***34.63***	***37.7***
SCS‡ (points)	***-0.3***	***1.3***	***0.5***	***1.3***	***0.71***	***1.1***
EQ-5D index (points)	0.098	0.075	***-0.047***	***0.084***	0.099	-0.037
EQ-5D VAS (points)	***3.4***	***3.9***	***-3.3***	***-10.0***	1.3	-0.6

† sample size for "much better" group for TUG & 2MWT in TI2,

‡ sample size for "much better" group for SCS in TI3,

‡‡ sample size for "much better" group for TUG & 2MWT in TI3

*p<0.05, **p<0.001

6.4.10.3 Hypothesis 3

"The outcome measures would correctly identify individuals who had and who had not undergone an important change 80% of the time."

This hypothesis can be tested by reviewing the ROC curve analysis. As can be seen in Table 6.16 presented earlier, the AUC was not above 0.8 for any of the outcome measures. The conclusion reached is that none of the outcome measures were able to correctly identify the participants who had and who had not undergone an important change 80% of the time. Results of over 70%, representing moderate accuracy, were obtained by the SIGAM, TUG, 2MWT and EQ-5D-5L for the index score, in TI2.

6.5 Discussion

6.5.1 Study focus and setting

The focus for this study was estimating responsiveness parameters, for outcome measures of physical function that are used regularly during the prosthetic rehabilitation of pwLLAs. The study was set during the early phase of prosthetic rehabilitation, when changes in the physical ability and function of these amputees are to be expected. It was therefore likely that changes would be seen in those outcome measures of the constructs under investigation. The gait training, strengthening programme or prosthetic componentry prescribed to participants during this early rehabilitation period are not described as they are not the focus of this study and therefore will not be discussed in any detail. Thirty unilateral adult pwLLAs were followed in this study, from approximately one week after they were fitted with their first prosthetic limb, until their MDT clinic appointment at six weeks after their hospital discharge. Distribution-based parameters of responsiveness (such as repeated measures ANOVA and effect size) were calculated for the SIGAM, LCI-5, TUG, 2MWT, SCS and the EQ-5D-5L. In addition, anchor-based parameters of responsiveness (MCID) were also calculated using the amputee's perceived change in their physical abilities.

During the early prosthetic rehabilitation period immediately after the amputee takes delivery of their first prosthesis, the physical function of a pwLLA is likely to improve, especially with regard to their mobility. By collecting data at three data time points during this period, it was possible to investigate if there were any more pronounced changes during the phase immediately after delivery of the prosthesis (TI1, duration approximately two weeks) compared to the phase following discharge from hospital (TI2, duration approximately six weeks). While this was not an explicit research question, it was noted that there was a superior improvement noted in TI1 compared to TI2, for all outcome measures except the SCS and the EQ-5D-VAS scores when considering the effect sizes derived from the change scores over each time interval. Only the LCI-5, TUG and 2MWT test results showed significant differences in both time periods (TI1 and TI2).

This, apparently rapid, improvement in mobility and functional activities is not unexpected soon after delivery of the prosthetic limb. The amputee is likely to be receiving daily gait training and physiotherapy during this in-patient period of

rehabilitation which will improve their mobility and function. If there are no issues with wound healing around the residuum and pain is well controlled then mobilising on a prosthesis will proceed at a pace dictated by the amputee. Demet et al (2003) postulated a link between energy levels and quality of life, in a study looking at the Nottingham Health Profile with pwLLAs. However, others highlighted the need to investigate the psychological adjustment of an amputee in the immediate post-surgery phase more (Horgan and MacLachlan 2004). It is interesting to note, that the pwLLA respondents to the survey reported in Chapter 3, put "energy, drive and motivation" in their top three factors that influenced a positive prosthetic outcome. This concurred with "determination to walk" and "motivation" that were considered important predictors in the rehabilitation process by prosthetic users and health professionals (Schaffalitzky et al. 2012). Pain levels were not collected in this study but it is expected that physical function and mobility will be adversely affected by increasing levels of pain and this would be reflected in the data obtained from the outcome measures. "Severity of pain" also emerged as an important predictor of prosthetic limb use by both prosthetic users and health professionals who participated in the Delphi Study carried out by Schaffalitzky (2012).

Following discharge from hospital, the amputee will be required to be self-motivated to continue their programme of exercises and gait practice to achieve his /her functional goals. However, it is possible that the rate of any improvement towards these goals may slow, or even level off, as potential everyday difficulties become a reality when the amputee goes home. This realisation, coupled with the absence of a daily therapy routine with their fellow patients and the health professionals to oversee the therapy programme, may contribute to a slowing of any functional improvements or in some cases, a decline. Such variations in performance were seen in a study by van Twillert et al (2014) where the physical performance declined relative to baseline, in nine out of thirteen pwLLAs who were followed at three and six months post-hospital discharge (van Twillert et al. 2014). The same level of deterioration was not seen in this current study. However the participants were only followed up until six weeks post-discharge, so it is possible that the decline in performance noted by van Twillert was not captured.

6.5.2 Perceived vs observed change

The additional data collected from participants via the ACQ at SV2 and SV3 provided an anchor of improvement perceived by the patient and was considered an indicator of the importance of the change. Construction of the anchor question is considered of fundamental importance to the design of a competent scale and should be constructed to align with the domain of interest (Kamper 2009), in this case, physical function. A time scale must also be included. By asking the specific question "how do you rate your current physical ability to perform everyday tasks is today, compared to when you first got your artificial leg" in the ACQ in this current study, it was anticipated that an indication would be gathered of how the pwLLA felt they had changed with specific reference to their physical ability to undertake everyday tasks. Despite the careful choice of wording in the change question, it is recognised that with only one anchor question but data from six outcome measures, all dealing with physical function but with slightly different constructs, there may be some issues with the alignment of the ACQ question. As the SCS and the EQ-5D-5L measure socket comfort and QoL, respectively, the wording of the anchor question may not be optimal for them. There will be more discussion of the validity of the ACQ for each outcome measures when they are presented individually, later in the discussion.

The majority of the pwLLAs in this study felt their "physical ability to undertake everyday tasks" was "much better" according to the ACQ scores. However, this did not always correspond to an improvement recorded by the outcome measures as evidenced by the poor correlations between the ACQ scores and differences recorded in each time period, (see Table 6.16). There was no relationship seen between the ACQ scores and either of the two observed measures (TUG and 2MWT) in any of the time periods. This lack of correlation between perceived and actual performance has also been seen in a review by Kannenberg et al (2014) when performance-based outcome measures such as the 2MWT and the Hill Assessment Index demonstrated improvements in patients using micro-processor knee joints but self-reported measure results showed a lack of improvement (Kannenberg et al. 2014).

There is some suggestion of recall bias with this method of collecting patient-perception of change (Kamper 2009, Norman et al. 1997). The exact phrasing of

the anchor question would appear to be key to the determination of the relationship between responses to the anchor question and changes in outcome measures collected at the same time; and any errors of measurement in the change questionnaire and any outcome measure being collected at the same time, are assumed to be unrelated. However, any influences on the patient in their current state will affect their recall of their previous state (Ross 1989). This is likely when the current state is being reported by both the patient or measured by an observer.

While there may be concerns about whether both PROMS and observed measures can effectively detect changes in performance, the use of PROMs is being heralded as having the potential to transform healthcare by some. Black (2013) postulates in his editorial analysis in the British Medical Journal, that PROMs, both disease specific, e,g, the Oxford Hip Score, and generic measures like EQ-5D, will not only "help patients and clinicians make better decisions, but they can also enable comparisons of [health care] providers' performances to stimulate improvements in services" (Black 2013). A discussion around reporting perceived versus actual performance and the use of both patient-report and observed outcome measures will be presented in more depth in the final discussion chapter (chapter 7).

6.5.3 Distribution-based parameters of responsiveness

Following repeated measures analysis of variance, statistically significant differences were noted across the full period of the study (median 74 days duration), for all outcome measures, except the SCS and EQ-5D-5L. The SCS and EQ-5D-5L were developed to measure the comfort of the socket and global health, respectively and this early rehabilitation period with the prosthesis is aimed primarily at functional mobility. So although there is an expectation that there will be some changes in both socket comfort and global health during this period, these changes may not be so pronounced. Post-hoc analysis showed significant differences during the rehabilitation period, immediately after limb-fitting in TI1 (median 12 days) for SIGAM, LCI-5, TUG and 2MWT. However statistically significant differences were only noted for the TUG and 2MWT in TI2, the post-hospital discharge period (median 64 days). Further discussion on these results will be presented later as each outcome measure is examined separately.

6.5.4 Anchor-based parameters of responsiveness

The detection of statistically significant changes in scores during a time when changes were expected, have led some authors to conclude that the outcome measure in question, is effective in detecting changes i.e. is responsive (Brooks et al. 2001, Panesar et al. 2001, Kohler et al. 2011). In a similar way, others have declared an outcome measure is responsive to changes in the participant's condition, by presenting effect sizes (Devlin et al. 2004, Franchignoni et al. 2004, Ryall et al. 2003). However, it is now recommended by many (Copay et al. 2007, Haley and Fragala-Pinkham 2006, Revicki et al. 2008) that in order to establish that an outcome measure is responsive to "clinically important" changes, these distribution-based methods are used to support anchor- or criterion-based MCID estimates as a responsiveness parameter. It is also recognised that before an outcome measure can be deemed responsive, for a given population and context, any changes detected must be relevant to that population and context (Copay et al. 2007, Haley and Fragala-Pinkham 2006, Terwee et al. 2012). The a priori hypotheses stated in this study, were formulated with respect to triangulation of results (statistical significance, magnitude and clinical relevance), and the sensitivity to detect change by each of the outcome measures have been assessed in relation to these. The MCID values for each outcome measure presented are also considered in relation to the validity of the ACQ and its accuracy across each of the time intervals.

In TI2, the improved group, who considered themselves "much better", consisted of 20 participants for the SIGAM, LCI-5, EQ-5D-index and EQ-5D-VAS, and 19 for the TUG, 2MWT and SCS. This was compared to only three who considered themselves only slightly better. Even for the other time intervals, the numbers for the "other" group are not large (i.e. 7, 8 and 9). Using a five point Likert change scale may have contributed to this imbalance, as the participants were not able to qualify how much they felt they had improved beyond the two choices; "slightly better" or "much better". End-aversion bias, also known as central tendency bias, may also render a five level scale equivalent to a three-level scale, and some authors argue that if a five-level scale is required seven options should be provided (Streiner et al. 2014). In a review of global rating scales, Kamper et al (2007) noted that scales with seven to eleven levels, balanced around zero, appeared to offer good patient acceptability with the ability to discriminate improvement or

deterioration adequately (Kamper 2009). Given that an improvement was to be expected in the time frame of this study, the five level scale used may have had insufficient levels describing “better” and will have compromised the ROC analysis and calculation of the MCID values, especially in TI2. Further discussion of this, and the impact of the ACQ validity testing results will be considered below and when the outcome measures are discussed individually.

6.5.5 Validity of the change questionnaire

Overall, the validity of the change questionnaire (ACQ) in this context was not confirmed. The accuracy of all the outcome measures in correctly identifying individuals who had and who had not undergone an important change based on the responses of the ACQ used in this study was also not good. Both the question asked and number of options available in the scale, may have been responsible for this. Improving the ACQ with more specific questioning about the different aspects of the constructs being measured in each individual outcome measure, and providing a higher number of response options may have improved its accuracy. However, calculation of MCID values remains difficult even when a gold standard anchor is used. In a study with stroke survivors, MCID values were calculated for gait speed in stroke survivors during the first 60 days post-stroke (Tilson et al. 2010). A well-established global criterion measure was used, the modified Rankin Scale (mRS). The mRS assesses changes across five levels and is considered a gold standard anchor to identify minimal clinically important changes in disability. A score is assigned by an assessor after a standardised three to four hour assessment. However, the authors still found the precision of the mRS was less than 80%, and they concluded that the anchor did not correlate directly with gait speed. The topic of how the accuracy of specific change questionnaires, for example the mRS versus global change questionnaires such as the GRC, impacts on ROC analysis and the calculation of MCID values, will be explored more fully in chapter 7 (final discussion).

A lack of confirmation of the validity of the ACQ challenges its use in the calculation of MCID values for each of the outcome measures and associated estimates of responsiveness. Evidence for the responsiveness of the outcome measures is available, from the results derived through the distribution-based methods, that some of the outcome measures are sensitive to changes in the pwLLA’s condition in

this early rehabilitation period. However, without support from the anchor-based derived results the clinical importance of any changes cannot be confirmed.

6.5.6 Specific outcome measures

For all outcome measures, three sets of results will be discussed. Firstly, responsiveness parameters derived from distribution-based methods, i.e. ANOVA and effect sizes. Secondly, results from an anchor based-method using the ACQ to derive MCID values will be presented. The accuracy of using the ACQ to derive these values for each outcome measure will also be considered. Finally the validity of using the ACQ to act as an appropriate anchor in this context will be discussed.
All results will be discussed with reference to the relevant literature,

6.5.6.1 Patient Reported Outcome Measures (PROMs) measuring functional mobility

6.5.6.1.1 SIGAM

The SIGAM grades are assigned by following an algorithm after a series of questions on functional mobility have been answered (Ryall et al 2003a). The purpose of a SIGAM grade is therefore to describe functional levels of mobility and the ACQ question appears to be well suited to the construct being measured by the SIGAM.

Main ANOVA and post-hoc analysis showed statistically significant group mean differences and large effect sizes across the whole study period (TI3) and the immediate limb-fitting period (TI1), demonstrating that the SIGAM was sensitive to changes within these time frames. However, the group mean differences between SV2 and SV3 (TI2) were not seen to be significant and the effect size was only small in this period after hospital discharge.

Despite the moderate accuracy seen for the ACQ from the ROC analysis for TI2, the MCID values obtained for all the outcome measures during this time interval may have to be discounted because of the low numbers in the “negative actual state” group (n=3). There was low accuracy for the ACQ in both the immediate limb-fitting period (TI1) and across the whole study period (TI3), but results obtained from the distribution-based methods did provide better support for the MCID values obtained

for the SIGAM. They were 2.5 (which translates into 3 grades) and 4.5 (5), for both TI1 and TI3 respectively.

The results from the hypothesis testing for this time period were mixed for both T1 and TI3. However, due to the limitations of the change questionnaire (described above) and the small sample numbers in each group, it is not possible to suggest that the hypotheses may be proven or confirmed, and so the results are discussed in terms of whether a hypothesis is likely to be proven or not. The indications were that hypothesis 1 (participants with higher ACQ scores will show a greater improvement) was unlikely to be confirmed, as there was very low, or no correlation between the ACQ scores and differences in all time intervals. The second hypothesis (the size of the change would be larger in those who perceived that their physical function was much better (ACQ=5) compared to those who did not) did show positive results for both TI1 and TI3.. Finally, hypothesis 3 (outcome measures would correctly identify individuals who had and who had not undergone an important change 80% of the time) was tested using the ROC curve. In TI1 and TI3 there was only slightly better than chance results with just over 50% for both Although there was moderate accuracy (just over 70%) for TI2, though as discussed earlier this result is likely to be compromised due to the very small numbers in the "negative actual state" or "other" group.

The validation results for the ACQ in TI3, across the whole study period demonstrated a poor performance in relation to the SIGAM, with only two out of the five criteria met: a positive correlation between the ACQ and SV1 (baseline) and a difference of >0.2 in correlations between SV1 and SV3. This means that the MCID value of 5 grades (total 11), calculated across the whole study period should be treated with caution. Similar validation results in TI1 (two out of the five validity criteria met) mean that the MCID value of 3, for the very early in-patient rehabilitation period, is also presented with some caution. The MDC that was presented for the SIGAM in chapter 5 was 1 grade, so both MCID values presented here are outside the measurement error value for this outcome measure. It should be noted that, although they were all pwLLAs, the samples on which the MDC and MCID values were calculated were not the same. This means that the results should be treated with caution and considered in this context, but it is still likely that both the MCID values are realistic in this rehabilitation setting.

It should be noted that the original six grades presented in Ryall et al (2003a) describe a progression of increasing independence and functional locomotor ability from limb abandoned/cosmetic only (A), through to normal or near normal walking (F). However, within the scoring system there are 11 levels of grading that take into account the increasing levels of walking support from for grades C and D. These 11 levels were assigned in this study and therefore the MCID values of 3 or 5 grades should be considered in relation to the increased number of levels. Using the 11 grade scale, a SIGAM score would change by one grade each time a walking aid was changed. With a reduction in the dependence on the type of walking aids noted between every visit in this study, a change of 3 to 5 grades may be considered reasonable in the context of this early rehabilitation period. Though it should also be noted, that moving from indoor to outdoor walking is equivalent to a change of 4 grades on this 11 grade scale. Therefore, it is would be advisable to carry out a Rasch analysis on the 11 grade scale to confirm the true interval nature of the scale, before these MCID levels are recommended.

A very large effect size of 10.6 has been presented for the SIGAM in a previous study also undertaken during rehabilitation following delivery of a first prosthesis (Ryall et al 2003a). Twenty five patients completed the SIGAM on 38 occasions throughout their rehabilitation period, however the time between assessments varied between 3 weeks and six months post-delivery of their prosthesis. The authors concluded that the effect size demonstrated that the SIGAM was sensitive to changes over time, though it was not clear how the effect size was calculated (Ryall et al 2003a). A timed walk was undertaken at the same time as the assessment visits and the results were used to demonstrate the construct validity of the SIGAM in that study. However, correlations of changes in the results of the timed walks with those of the SIGAM grades to confirm improvement were not presented. Nor were any other parameters of responsiveness and it was therefore, not possible to establish whether the SIGAM was a responsive outcome measure from that study. This evidence was rated unknown in the systematic review reported in Chapter 4.

6.5.6.1.2 LCI-5

A large effect size was demonstrated and significant differences were seen between the group means for the LCI-5 scores across the whole period of the study,

providing evidence for the responsiveness of the LIC-5 from distribution-based parameters. There was a similar pattern, but with medium effect sizes in both TI1 and TI2.

The LCI-5 demonstrated moderate accuracy when used with the ACQ across all time intervals. The MCID values obtained for both TI1 and TI2, 8.5 and -1.5 respectively (translating into 9 points for TI and 2 points for T2), are supported by the distribution-based results. These changes appear to be reasonable and might be expected as the pwLLA's mobility improves and the dependence on stand-by support of another person and on their walking aids decreases. The support required from other people by the participant was not assessed, but a reduction in the dependence on walking aids was noted between every visit. It might therefore be reasonable to assume that there could be an average swing of 14 points on the LCI-5 scale if the amputee requires assistance with all the basic tasks at his/her first visit and considers himself/herself able to perform the same tasks independently with walking aids at the final visit. Potential for greater change scores in patients who have a greater level of disability and consequently lower baseline scores has been reported previously (Stratford et al. 1998). However, a change of 21 points on the LCI-5 (MCID = 21 at TI3), may be unlikely, in this early rehabilitation period. An MDC of 4 points was presented in chapter 5 for the LCI-5, and as all the MCID values calculated in this study for the LCI-5 are well outside that, they could all be calculated in this context.

The ACQ, again, demonstrated only low accuracy in its ability to detect those participants who had improved on the LCI-5. It was expected that the anchor question was most related to the constructs that the LCI-5, and the SIGAM, are measuring and therefore does question the appropriateness of the wording of change questionnaires. However, it is possible that a change in LCI-5 scores may occur, *before* the pwLLA perceives an improvement in his/her physical abilities. i.e. while the amputee recognises that they do not need the help of others and his/her dependence on walking aids lessens consequently lowering the LCI-5 score, they do not consider this to be a large improvement in physical function.

The greater effect sizes and group mean differences in the much improved group (ACQ 5) compared to the other group (ACQ 1-4) for all time intervals is suggestive

of a positive outcome for hypothesis 2. , hypothesis 1 was not confirmed and hypothesis 3 demonstrated only slightly better than chance results. The validation results for the ACQ in all time intervals demonstrated a very poor performance in relation to the LCI-5. Only one out of the five criteria were met in TI1 and TI2 and only two out of five in across the whole study period in TI3. This, in addition to the poor accuracy of the ACQ in detecting important change, means that all MCID values should be treated with caution.

Franchignoni et al (2004) presented an effect size of 1.4 in 2004 for a sample of 50 pwLLAs (median age 51) studied during prosthetic training. The period over which the changes were measured averaged 36 days, which was slightly shorter than the post-hospital phase in this study. However, this effect size of 1.4 was comparable to those recorded for the whole period in this study (i.e.1.15). No other parameters of responsiveness were presented and no further comment was made about the responsiveness properties of the LCI-5 by the authors. The evidence for this LCI-5 study Franchignoni et al (2004) was rated unknown in the systematic review reported in Chapter 4.

6.5.6.1.3 SCS

With no significant differences and small effect sizes recorded across all the time periods, there is no evidence that the SCS is a responsive outcome measure when using distribution-based methods.

There is also little support for the robustness of any of the MCID values as none of the experimental hypotheses were likely to be upheld. In addition, the validity of the ACQ was poorly demonstrated in all time intervals. There is also no evidence that the SCS is a responsive outcome measure in any of the time periods when using a patient-anchor to detect changes in the physical ability of pwLLAs to undertake everyday tasks. This may be because of the poor relationship with the question in the change questionnaire. It may also be argued that socket comfort, or discomfort, does not affect physical function as much as was thought, or that the SCS is not sensitive enough to measure the subtle changes in physical ability of the pwLLA that are affected by changes in socket comfort. It is relatively easy to see when socket discomfort reaches a level that stops an amputee from walking altogether, but when the discomfort impacts only on the quality, and not the quantity, of their walking then

more subtle measures may be required to pick up these changes. However, given rapid changes can occur in the socket / residuum interface on an hourly, let alone daily or weekly basis, estimating MCID values may be too challenging for this outcome measure. Hanspal et al (2003) did demonstrate statistically significant differences using Wilcoxon analysis in SCS after socket adjustments, though no other parameters of responsiveness have been presented since that date by any authors.

6.5.6.1.4 EQ-5D-5L: Index score and VAS

There is limited evidence that the EQ-5D-5l is responsive to perceived changes in physical function in pwLLAs as shown by low effect sizes and lack of statistical significance when both the index scores and VAS are scrutinised. No significant differences and small effect sizes were presented for the EQ-5D-VAS score in all of the time periods. It is a similar pattern for the index scores except a medium effect size was noted in TI1. Although the majority of effect sizes for both the index scores and VAS are small, there appears to be an anecdotal trend (no significance calculated) indicating deterioration in the pwLLA's perceived global health status following discharge from hospital.

Despite the medium effect size seen in TI1 for the EQ-5D-index score, there is little support for the precision of an MCID value of 0.108 because of the low accuracy demonstrated with an AUC of 0.552. In addition, the validity testing was weak in all time intervals and therefore, there is also poor evidence that the EQ-5D-index score is responsive in any of the time periods when using a patient-anchor to detect changes in the physical ability of pwLLAs to undertake everyday tasks.

None of the hypotheses were likely to be confirmed for either the EQ-5D-index or VAS scores, though there was moderate accuracy for the ACQ was seen in TI2 for VAS scores. Despite the moderate accuracy there is little support for any of the MCID values obtained for the VAS, with the validity also testing poorly in all time intervals.

The poor relationship between the ACQ question and the construct measured by the EQ-5D-5L (global health) may be the cause of these disappointing results. In a similar way to socket comfort, it could be argued that global health, does not affect physical function in pwLLAs, as much as was thought, or that the level of physical

function does not affect the global health of pwLLAs as much as we thought. It may also be that any changes in the global health of the amputee is masked in the early rehabilitation period. It is possible that the EQ-5D-5L is not sensitive enough to detect subtle changes in physical ability of the pwLLA, that are affected by changes in the global health outcomes it measures, though the index score did register an effect size of 0.62 in TI1. In the review by Horgan and MacLachlan (2004), it is argued that the period following discharge from a regular rehabilitation centre has not been well documented with regard to psychosocial adjustment. The amputee has not yet entered the "long term adaptation phase" where they must adapt to the reality of their new life. The relative short time-frame of this study may therefore not capture this realisation.

There are no previous responsiveness results for the EQ-5D reported for a pwLLA population, but a review of minimal important difference (MID) values, similar to MCID, have been reported with other populations. In eight studies that looked at the use of EQ-5D (3-level) in patient groups with a range of health conditions, a mean MID value of 0.074 was reported (range -0.011 to 0.140) (Walters and Brazier 2005). The MID value in each study was calculated by taking the mean change on the EQ-5D score of those patients who were considered as having experienced a change, as measured by Question 2 in the SF-36 (5 level rating scale of change). The authors noted the difficulty of establishing responsiveness parameters for utility questionnaires, such as EQ-5D, because of their complex nature "reflecting both preferences as well as status on health dimension". They felt that more research was required to fully understand the psychometric properties of these complex measures in other patient groups and populations. It was also interesting to note, that Walter and Brazier (2005) felt that unlike the EQ-5D, other (types of) outcome measures have been studied "extensively" with regard to distribution and anchor-based methods of establishing responsiveness. However, this "extensive" work has not been seen in outcome measures used with pwLLAs.

6.5.6.2 Observed measures of mobility, balance and endurance

6.5.6.2.1 TUG mobility / balance

Statistically significant differences were seen between all the study visits and associated medium (TI1 and TI2) and large (TI3) effect sizes were also recorded;

thus, demonstrating the ability of the TUG to detect changes in the physical ability of pwLLAs during the early rehabilitation period.

Despite these distribution-based results in TI1 and TI2 the MCID values calculated for these time periods are not well supported by the results from the ROC curve analysis. In TI1, the MCID (0.3s) was less than the measurement error value of 1.8s, for the TUG presented in chapter 5. This together with the accuracy, that was no more than chance would mean that the MCID value should be disregarded. As previously stated for the SIGAM, despite moderate accuracy seen for the ACQ from the ROC analysis for the TUG in TI2, the MCID values obtained may have to be discounted because of the low numbers in the "negative actual state" group (n=3). The MCID of 5.5s for TI3 is presented with some caution because of the very low accuracy (on the cusp of a chance result) seen from the ROC analysis. However, it is greater than the measurement error value of 1.8s and thus detectable. It is also a reasonable amount of improvement that you may expect to see in the TUG during the early rehabilitation period, considering the 11 week period covering TI3 (mean 84 days).

There was little relationship demonstrated between the ACQ and the change scores in any of the time intervals. Moderate accuracy for the ACQ was found for TI2, but it did not reach 80%. Therefore the results suggest that there was little support for the any of the hypotheses. Overall this means there is little support for any of the MCID values that were calculated using the ROC analysis. In addition, the ACQ validity testing demonstrated only two out of the five criteria, at best, in any of the time intervals.

There are several possibilities for the lack of accuracy and validity results: Firstly, the ACQ question may be too general to relate to the specific constructs of balance and mobility that the TUG is measuring. While a large number of pwLLAs did perceive an improvement in how they were performing everyday tasks, it is possible they did not relate this to how they were performing in relation to balance and mobility as assessed in a timed test. As discussed previously, formulating a specific question for each outcome measure may have allowed the changes perceived, to be related more closely to the specific constructs measured by the TUG. For example the question could have been "how is your ability to get out of a chair, walk a short

distance turn around and sit down again, today, compared to…." However, it could be argued, that it is not relevant to focus on each single task, when there are so many activities of daily living (ADL) that an individual amputee will undertake. Secondly, as discussed above it is also possible the number of options given on the ACQ was too small. Finally; there is the possibility that the answer to the anchoring question was influenced by pwLLA's current state or perception on his/her current state thus affecting their recall (Ross 1989). However, previous times were not shared with the participant or the assessor at the time of the current visit and the ACQ was always acquired first at the study visit to avoid any overlay of performance bias.

There are no published results for the responsiveness of the TUG in pwLLAs but Rushton et al (2014) did report an MCID value for the L-test for pwLLAs in a pilot study (n=33) that had been completed in 2014. The L-test is closely related to the TUG, with similar start and finish points, i.e. sitting in a chair, but the patient is asked to stand up, walk to a cone 3m away, turn left (or right) to a second cone, and then retrace his/her steps back to the chair. Similar to the TUG, patients are timed doing these manoeuvres. Participants in the Rushton study completed the L-test at baseline and then after variable lengths of rehabilitation (1.5 to 11.5 months and not detailed). A global rating change (GRC) scale was collected at the follow-up visit, which rated the participant's perception of a change in their condition using a 7 point scale. Participants were asked about their "…ability to get up and walk with your prosthesis ……." if it had improved they were asked to rate the improvement using a 7-point Likert scale, from 0 = almost the same, to 6 = a very great deal better. If they had deteriorated there was a similar 7-point scale using a "worse" perspective. Thus they used a 13-point scale. Two a prori hypotheses were postulated in a similar way to those in this current study: Hypothesis 1 - individuals who rated themselves as having greater GRC scores would have greater changes in L-test scores and; Hypothesis 2 - the L-test would correctly identify pwLLAs who have and have not undergone important changes in >80% of cases. Hypothesis 1 was confirmed, but hypothesis 2 was not supported. The L-test did identify patients having a perceived an important change, (i.e. those who score ≥ 5 on the positive or improved side of the GRC), but the GRC scale did not score highly on accuracy. The authors speculated that this was possibly due to recall bias related to the length of the rehabilitation intervention as some participants were not retested until nearly a

year after their baseline visit. As discussed previously, the current state may influence recall of a previous state (Ross 1989). Validity of the GRC scale was also tested by the authors, Rushton et al (2014), and it was found not to be a valid tool to assess important change in "the ability of an individual with a pwLLA to get up and walk with a prosthesis". Despite this, the authors went on to quote an MCID value of 4.5s that was calculated using the GRC scores in a ROC analysis (AUC 0.67). They did say that their results must be interpreted with caution in light of these low accuracy results. Therefore, despite attempting to use a specifically framed question for the L-test in their anchor questionnaire and using a 13 point scale, Rushton et al (2014), also found the predictive value was also less than 80%, though did comment that it was "better than chance alone".

The merits, or otherwise, of using a 13 point scale, such as the one used in the Rushton et al (2014) study, versus the 5 point scale used in this current study, will be explored further in the final discussion chapter.

6.5.6.2.2 2MWT mobility / endurance

The results obtained for the 2MWT presented a similar scenario to the TUG with regard to the distribution-based results. Statistically significant differences between group means and medium (TI1 and TI2) and large (TI3) effect sizes were recorded across all the study time intervals, thus demonstrating the ability of the 2MWT to detect changes in the performance capabilities of pwLLAs. However, despite these significant differences, and medium and large effect sizes recorded, the MCID values calculated for the 2MWT are not well supported by the results from the ROC curve analysis in any of the time periods. The accuracy of the ACQ to detect changes was no more than chance, according to the ROC analysis, in either time interval. Therefore despite the MCID values of 57.9m (TI1) and 55.4m (TI3) being well outside the MDC value (18.9m) presented in chapter 5, the results must be viewed with caution in view of these accuracy results.

The greater significant differences in the "improved" group compared to the "not-improved" group for both TI2 and TI3 were the only results that may indicate that one of the hypotheses (hypothesis 2) might be confirmed. The accuracy of the ACQ to detect these changes did not reach 80%, and with only low accuracy recorded for TI2, and no better than chance for both TI1 and TI3, it was unlikely that hypothesis 3

would be proved. Taking into account the limitations described earlier, it appears that there is little evidence that the ACQ is a valid tool as a patient-anchor to detect changes in the 2MWT in any of the time periods. This is because of the validity testing demonstrating only one out of the five criteria, at best, in any of the time intervals, and none of the hypotheses likely to be accepted.

It may be argued, in the same way as for the TUG, that by changing the question on the ACQ, this time to "how do you feel your ability to walk continuously is today, compared to…….", a more accurate picture of the responsiveness of the 2MWT in this population, would have been presented. This would be a recommendation for any future study that aims to assess the responsiveness of the 2MWT in any population.

Brooks et al (2001) demonstrated significant differences in distances walked by pwLLAs before and after rehabilitation, using repeated measures ANOVA. No other responsiveness parameters have been presented for a population of pwLLA. However, MCID values have been presented for the six minute walk test (6MWT) in COPD patients (Polkey et al. 2013). Data from lung function test producing FEV1 (1 minute forced expiratory volume) values and 6MWT times were collected (amongst other clinical data) from 1,847 COPD patients in a 3-year, multicentre, longitudinal, prospective study. At 3 and 6 months, and every 6 months thereafter, health status was measured using the COPD-specific St. George's Respiratory Questionnaire (SGRQ-C). The authors attempted to use this health status questionnaire in an anchor-based method to calculate MCID values for the 6MWT but the correlations were so poor between the SGRQ-C and change scores for the 6MWT that the method was abandoned. Instead, the MCID values for 6MWT were calculated retrospectively using mortality as an index and FEV1 values as the indicator of deteriorating cut-off points, i.e. 6MWT distances can then be related to respiratory function to give an indication of any deterioration, rather than use perceived health status (Polkey et al. 2013).

6.5.7 Grouped outcome measures: PROMs related to observed

The results for the individual outcome measures have been considered separately. An analysis of the correlations between changes in observed outcome measures and those in PROMs, was not undertaken. Any advantages of collecting both

PROMs and observed outcome measures will be explored further in the final discussion chapter (Chapter 7).

6.6 Limitations

The study was of a longitudinal cohort design, and patients were recruited on a sequential basis as they fitted the criteria and consented to be included. To ensure recruitment of the sample size required by the calculations in a reasonable timescale, a pragmatic decision was made to include both TT and TF levels of amputations. It was recognised that recruitment of both TF and TT amputees was a potential source of variation. However, as participants are asked to consider their own progress in the anchor questionnaire and are therefore their own control, within-subject analyses were undertaken and not between-subjects. It would have been advantageous to consider the results by age, aetiology or amputation level, but, the sample size was too small to do an in-depth analysis on sub-groups. Additionally, the sample size in this study would also render the evidence collected in this study to be rated as "unknown" according to the COSMIN checklist.

It could also be argued that the sample size was also too small to be considered for ROC analysis as Hanczar et al (2010) recommend that ROC curves must be used with extreme caution unless one has a very large sample (>1000) (Hanczar et al. 2010). Though even with a large sample size, the differences between true and estimated error can be considerable and they recommend that unless the lower 95% confidence for the AUC is greater than 0.5 (i.e. greater than chance) no conclusions can be reached.

As discussed above, only one "general physical function" question was asked in the ACQ, it may have been more appropriate if specific questions relating to different constructs of each outcome measure had been included.

The ACQ used a five-level response questionnaire which, it can be argued, may not detect the small incremental changes that a larger number of levels would. Larger numbers in each group i.e. "better" "the same" or "worse", would ensure more robust ROC curve analysis with less tied results as was seen in this study. However, when calculating MCID values using instruments with these higher numbers of levels,

there is a risk of arbitrary decisions being made when deciding which groups to combine for the analysis (Copay et al. 2007). One way round this could be to present multiple MCID values for the range of responses (Stratford et al. 1998), however larger sample sizes would be required.

There are many other factors that will influence the change scores throughout the course of this study, including the type and duration of gait training and therapy that each participant receives. Details of these were not recorded as it was the impact of "total package" of rehabilitation as required by each individual that was of interest, and not the impact of specific aspects of the rehabilitation process. The reduction of the participants' dependence on walking aids was however, noted throughout the study. It is recognised that this will have had an influence, not only on the times and distances recorded for the TUG and 2MWT, but also on the SIGAM and LCI-5 scores. However, it is not known by how much. The use (or not) of walking aids was not accounted for in the data analysis as a confounding factor because of the superficial nature of the data collection. Consideration of such influences in any future study on the use of outcome measures, would be of value in interpreting the context of the recovery of the pwLLA. Increasing independence has been shown to help improve the participant's confidence and self-efficacy (Schaffalitzky et al. 2011), but again it is not known by how much this will have influenced the times, distances and scores recorded in this study.

6.7 Conclusions

All outcome measures, except the SCS and EQ-5D-5L, were responsive to changes across the whole study period, as evidenced from the ANOVA and effect size results.

The LCI-5 and 2MWT recorded larger effect sizes in the in-patient period when compared to the post-hospital discharge period. The effect sizes were similar in both the in-patient and post-hospital discharge periods for the TUG.

Validity testing for the ACQ was confirmed (by achieving more than 50% of the criteria) for only one outcome measure, the SIGAM in TI1.

Limitations of the validity of the change questionnaire used and the small sample numbers in each group, deemed it impossible to positively claim that the hypotheses were proven or confirmed. However, the following may be likely outcomes, as suggested by the results obtained.

Hypothesis 1. i.e. participants who rated themselves as having undergone a greater improvement in their physical function (i.e. higher ACQ scores) did not show a greater improvement in the six outcome measures over the same time interval and was therefore likely to be disproved for all the outcome measures..

Hypothesis 2. i.e. the magnitude of the change in the six outcome measures would be larger in those who perceived that their physical function was much better (ACQ= 5) compared to who did not (ACQ 1-4), was likely to be upheld for SIGAM, LCI-5 and TUG in TI1, for LCI-5 and SCS in TI2 and for SIGAM, LCI-5, EQ-5D-index and EQ-5D-VAS in TI3.

Hypothesis 3 was unlikely to be confirmed for any of the outcome measures, i.e. none of the outcome measures correctly identified individuals who had and who had not undergone an important change 80% of the time. However, it was noted that SIGAM, TUG and EQ-5D-VAS achieved over 70% accuracy in TI2.

MCID values were established for all outcome measures during all time intervals. However, results of the hypothesis testing did not support the accuracy of any of the outcome measures, when used with the ACQ as a measure of patient perception of change. The lack of confirmation of the validity for the ACQ also weakened the support for the MCID values. Therefore, the values that have been presented should be used with caution and may have limited clinical use.

7 Chapter 7 Final Discussion

7.1 Purpose of chapter

The purpose of this final chapter is to consider and synthesise the findings from the empirical studies, i.e. study III, the repeatability study (chapter 5) and study IV, the longitudinal cohort study (chapter 6). This final discussion will take into account the context of: the current usage of outcome measures in prosthetic rehabilitation in the UK, obtained from the survey (chapter 3); and the current supporting evidence for their psychometric properties as established by the systematic review of the literature (chapter 4). Thus the original parts of the thesis will be summarised and presented, in order to develop the current knowledge and understanding of outcome measures of physical function as they are used with pwLLAs (adults) during prosthetic rehabilitation. Recommendations for clinicians and researchers regarding the application of the most commonly used outcome measures to both, the individual and groups will also be presented, with the limitations of the studies taken into account.

7.2 Introduction

The research topic of this PhD was prompted by an in interest in measuring the impact of prosthetic componentry on the lives of pwLLAs. With ever increasing healthcare costs it is vital that clinicians, and also researchers, involved in the field of prosthetic rehabilitation, understand how to measure the impact of prosthetic componentry on their patients. Is one knee “better” than another? What does “better” mean? If, for example, this means, “better mobility” then how is mobility measured?

To evaluate the effectiveness of any intervention, the outcome measure being used to assess the impact must possess robust psychometric properties. Knowledge of an outcome measures’ psychometric properties is vital when assessing differences between one prosthetic component and another. The cost difference between two prosthetic knees can be thousands of pounds. However, information from appropriate outcome measures can objectively support the clinical decision of which

component to choose by detecting significant differences in, for example, the functional abilities of a pwLLA.

The background chapter (chapter 2) outlined the need for improved understanding of which outcome measures are regularly being used with pwLLAs. The survey (study I) conducted with clinical staff provided information in this area and helped chart the current practice in the UK. The current supporting evidence for outcome measures developed for use with pwLLAs, was outlined in the systematic review (study II). Thus, the central tenets of the repeatability study (study III) and the longitudinal cohort study (study IV) were formed because of the lack of specific evidence for measurement error and responsiveness for the outcome measures most commonly used by clinicians working in prosthetic rehabilitation.

An overview of the thesis, visually presenting the relationship between the research questions and how the gap in the current evidence was addressed, is presented in Figure 7.1.

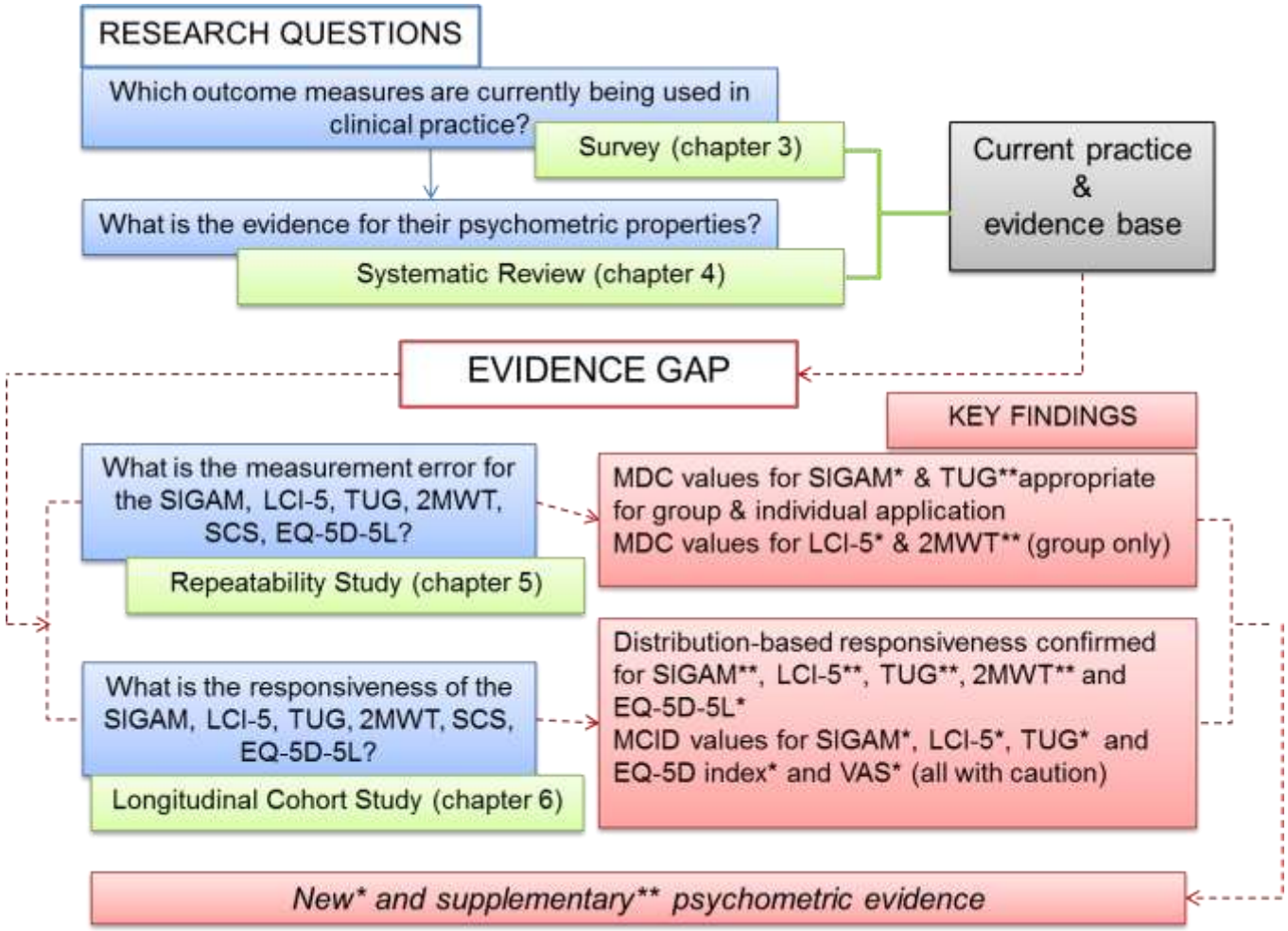


Figure 7-1 Final thesis overview

7.3 Outcome measurement in prosthetic clinical practice

7.3.1 Current practice

The findings from the survey showed that a wide variety of outcome measures are being used during prosthetic rehabilitation in the UK. Survey questionnaires were distributed via professional networks and sub-groups to AHPs working in prosthetic rehabilitation. They were distributed by e-mail to encourage onward distribution to multi-disciplinary colleagues, while the patient survey questionnaire was distributed to suitable pwLLAs by their prosthetist in the Edinburgh Prosthetic Centre. Forty-five AHPs and 12 pwLLAs responded. Thirty two of the 45 AHP respondents (71%) indicated they used outcome measures regularly, i.e. at least once a week. Physiotherapists that indicated they used a total of 14 different outcome measures from the list. Prosthetists indicated they used 8 different measures though they used two (the SIGAM and the SCS) most regularly. The occupational therapists used three, namely; the Barthel Index, COPM and FIM. This variety still indicates that there is a lack of consensus on which outcome measures to use. The five outcome measures used more often than any others were: the SIGAM, TUG, Timed Walk Tests (TWT), LCI / LCI-5 and SCS. This group of outcome measures all measure constructs in the physical domain, which may not be surprising given that the majority of respondents were physiotherapists. The breakdown of respondents, by professional group, was: 23 physiotherapists, 17 prosthetists and 5 occupational therapists.

Despite this bias towards outcome measures of physical function, “energy, drive and enthusiasm” was considered one of the top three most important factors to have a positive influence on prosthetic outcome. The other two factors also recognised by both the AHP and pwLLA respondents were “socket fit and comfort” and “balance and coordination”. While these results are from a very small sample of pwLLAs, and should be considered with caution, they do concur with the some of the findings by Schaffalitzky et al (2012) from their focus group work. They identified psychosocial factors, such as support from the family and MDT, as well as returning to work in a supportive environment were among those factors that are recognised as being facilitators to a positive outcome for the patient.

These psychosocial factors are not explicitly assessed in any of the five most commonly used outcome measures in current practice, though they may be

recorded in other ways in the patient's medical notes. It would be beneficial if outcome measures measuring different constructs were collected by a variety of health professionals in the MDT, as this could lead to a more comprehensive understanding of the progress (or lack of progress) of the patient during rehabilitation. However, consensus on which outcome measures to collect must be gained by the MDT. This approach, to bring together a comprehensive view of the patient's condition, performance and capabilities in the context of different clinical settings, is advocated by the COMET Initiative (Williamson et al. 2011). The initiative aims to bring together people interested in the development and application of agreed standardised sets of outcomes, known as "core outcome sets" (COS). The premise for this approach is that the COS should represent the *minimum* data that should be measured and reported in clinical trials of a specific condition (Macefield et al. 2014, Williamson et al. 2012). If the same data is collected from clinical trials conducted at single centres then data may be combined in meta-analyses more easily. Combining data in this way is one solution to small sample sizes and the difficulty in recruiting large numbers into rehabilitation studies that has already been noted.

Although the COS has an emphasis on clinical trials, they are also suitable for use in clinical audit and research, other than randomised controlled trials. Collecting COS data will make it easier for the results of trials to be compared, contrasted and combined as appropriate; but it is anticipated, by the COMET authors, that researchers could continue to explore other outcomes and not be confined to the COS alone. An example of a clinical area where a lot of work has been done developing this approach is Rheumatology with their OMERACT (Outcome Measures in Rheumatology Clinical Trials) initiative (Boers et al. 2014, Tugwell et al. 2007, Bautista-Molano et al. 2014), though many other clinical teams are working on developing and using COSs, (http://www.comet-initiative.org/studies/search/)

The survey results in Chapter 3 showed that while some outcome measures were favoured, some respondents are discontinuing the use of these same outcome measures. For example, the reason given for discontinuing two of the most commonly used outcome measures, the LCI-5 and SIGAM, were: "lack of time" and "no relevance to clinical practice". Clinicians are also still attempting to use their own "in-house" outcome measures. Anecdotally, from comments received in the

survey and also talking to colleagues working in prosthetic rehabilitation, it would appear clinicians prefer to use the outcome measure that they are most familiar with. This is usually the one that is easiest to implement (no costly equipment and/or doesn't take too long) and is the most acceptable to their patients, again usually the one that doesn't take too much time. Thus familiarity and clinical utility appear to be the most important factors deciding which outcome measures are used. Evidence from the literature does support this, with a lack of confidence in administering and analysing data from outcome measures, shown to be a barrier to their implementation (Jette et al. 2009, Gaunaurd et al. 2014).

The ease with which an outcome measure can be implemented must not be the only criteria by which a clinician makes a judgement on whether to use an outcome measure or not. Many guidelines and advice papers have been published for health professionals to assist them in their choice (Hammond 2000, Herbert et al. 2005, Corr and Siddons 2005, Cole et al. 2014, Young et al. 2015). These guidelines cover topics such as: establishing the outcome or construct to be assessed; understanding the psychometric properties that an outcome measure should have and; how to present the results. With ever increasing healthcare costs there is a continued emphasis on evaluating services and treatments to ensure value for money. Choosing the correct outcome measure with which to demonstrate the effectiveness of interventions has never been so critical. Therefore, choosing an outcome measure that has demonstrated validity, reliability and responsiveness in the population of interest, has to be as important for clinicians as its clinical utility because using an inappropriate outcome measure to demonstrate effectiveness will waste both their and their patients' time.

7.3.2 Current supporting evidence

Clinicians and researchers should be looking for evidence from high quality studies to reassure them of the validity, reliability and responsiveness of the outcome measure(s) they wish to use. However, chapter 4 showed that there is a lack of high quality studies reporting the psychometric properties of physical function outcome measures used with lower limb amputees. Ratings of the methodological quality of many studies were reduced because of weak description within the manuscripts and poor choices for statistical analysis, especially with regard to responsiveness. Consequently, there is limited evidence to inform the clinicians'

and researchers' choices. The systematic review reported in Chapter 4 concluded that while the LCI-5 had the strongest evidence to recommend its use for measuring functional mobility, the evidence was strong only for internal consistency (one study of excellent and two studies of good methodological quality presenting positive evidence) and structural validity (three studies of good quality). The evidence remained unknown for measurement error and responsiveness.

There is a lack of corroborating data (i.e. more than one study) for many of the outcome measures used in prosthetic rehabilitation, in spite of the many studies that have investigated and reported the measurement properties of these outcome measures. The systematic review (reported in Chapter 4) also showed a lack of studies reporting on responsiveness, and in particular the MCID values. The number of outcome measures that could be recommended to investigate the effectiveness of any intervention, based on its psychometric properties in this population was zero. There are several reasons for this lack of evidence: low methodology quality of the studies; small sample size; poor choice of analysis method and lack of corroborating studies to confirm findings. The limited choice of outcome measures, with robust psychometric properties, creates a problem when compiling credible evidence for the relative effectiveness of different prosthetic components (Kannenberg et al. 2014, Sawers and Hafner 2013). It also reduces the confidence with which clinicians and researchers can make an evidence-informed selection of outcome measures to assess progress (or lack of) of patients who have undergone a lower limb amputation.

Concern about the quality of prosthetics research evidence prompted a recent review of the literature (Hafner and Sawers 2016). The aim of the review was to identify if there were any common methodological quality issues that may be affecting the level of evidence reported, in this case specific to the evidence being provided on MCPKs. Common issues identified across the studies included: variable comparison conditions within studies, lack of blinding, small sample sizes, limited evidence of measurement reliability, participant attrition, and limited descriptions of participant selection criteria. The authors acknowledge that although the methodological quality of studies is improving there still needs to be more improvement in how the evidence is gathered and reported. They also felt that educating recipients of the research, i.e. clinicians and researchers, of the

importance of the characteristics that are inherent to prosthetic research was critical to improving the quality of the evidence required. The output of this thesis may help with this education.

Several standards are currently available to help with the reporting of research and clinical studies, including: Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) (Moher et al. 2016, Moher et al. 2016, Moher et al. 2009), Consolidated Standards for Reporting Trials (CONSORT) (Schulz et al. 2010), Transparent Reporting of Evaluations with Nonrandomised Designs (TREND) (Des Jarlais et al. 2004), and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) (von Elm et al. 2014). More details on these and other guidelines can be found on the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) website (http://www.equator-network.org/). While these standards and guidelines appear to have been responsible for improving the quality of the research evidence being published (Hopewell et al. 2010), it was noted that an editorial calling for the mandatory use of reporting guidelines that was simultaneously published in all participating publications, did not include any prosthetic and orthotics journals (Chan et al. 2014).

The standardised checklists presented above only examine the quality of studies that investigate the impact of treatments and interventions not those studies exploring the psychometric properties of outcome measures. Within these checklists the quality of any outcome measure used is only assessed with a cursory yes/no check, on whether the measure(s) used were “reliable” and/or “valid” and/or “responsive”. Therefore, there still remains a question over the integrity of the measurement properties of some outcome measures used in intervention studies, even if the methodological quality of these studies is high. The COSMIN checklist, as detailed in Chapter 4, is a specific checklist concerned with the quality of studies investigating the psychometric properties of outcome measures. It incorporates a multi-level grading system for each measurement property, thus giving detailed information on the quality of outcome measures to help support (or not) their use. A specific standardised checklist, such as COSMIN, investigating the psychometric properties of outcome measures, should be included in the list of those being recommended by journal editors. This would help to strengthen the quality of, and confidence in, the outcome measures being used in clinical practice and research;

and consequently strengthen any evidence from interventional studies that use them.

It was also seen in the systematic review, reported in chapter 4, that the methodological quality of studies was often compromised because of inappropriate statistical analysis. It would therefore be a recommendation from this thesis that in addition to guidelines on methodological quality requirements for intervention studies, that clear and understandable advice on the use of appropriate statistical analyses should also be made widely available. For example, the COSMIN checklist could be used as guidance when designing, as well as reporting, studies evaluating the psychometric properties of outcome measures in prosthetic rehabilitation. This will further improve the quality of the evidence, as will the type of collaborative work recommended by the COMET initiative (Williamson et al. 2012). Further discussion on the appropriateness of particular statistical analyses and their interpretation in studies investigating the measurement properties of outcome measures, will be undertaken later in this chapter.

7.3.3 Contribution to psychometric evidence

Further studies investigating the psychometric properties of outcome measures used with pwLLAs are recommended (section 4.8 (future work) in chapter 4). These studies are required to provide evidence, especially in relation to the reporting of measurement error, responsiveness and MCID values, to improve the utility and clinical applicability of outcome measures used in prosthetic rehabilitation. Although many studies in the systematic review reported on the reliability of outcome measures, the majority only reported indices of consistency but did not comment on agreement. Further, there was limited evidence of the measurement error values and responsiveness for outcome measures of physical function used with pwLLAs. These measurement properties are especially pertinent when evaluating the effectiveness of a treatment programme or prosthetic component. Therefore, the aims of studies III and IV, reported in chapters 5 and 6, were to establish estimates of reliability (including measurement error) and responsiveness for outcome measures regularly used by AHPs across the UK during prosthetic rehabilitation.

7.3.3.1 Consistency, agreement and measurement error

In study III (the repeatability study reported in chapter 5), estimated values for consistency, agreement and measurement error were calculated using a test re-test study design with an intervening period of 7 days between tests. The results were obtained from a sample of 12 pwLLAs who were established walkers and who were not attending therapy or undergone any prosthetic component changes in the previous three months. The stable condition of the participants and the strict repetitive nature of the administration of the tests attempted to reduce any systematic errors to a minimum. The MDC values for the outcome measures in study III were derived from the SEM for the study population and the results were presented to a confidence level of 95% (MDC_{95}). The key findings from this study are presented in Table 7.1.

Table 7.1 Key findings from study III

Research Questions (study/chapter)	**Key Findings**
Question 5: What are the reliability and measurement error parameters of outcome measures most regularly used by AHPs for the assessment of physical function during prosthetic rehabilitation? (Repeatability study - study III / chapter 5)	SIGAM: MDC=1 grade TUG: MDC=1.8s Both demonstrated excellent reliability (consistency ICC >0.9) and agreement with LoA of 1.2 grades and 3.4s (respectively) and can therefore be recommended for use at the individual level. 2MWT: MDC=18.9m LCI-5: MDC=4 points Both demonstrated excellent consistency (ICC >0.9) but questionably wide LoA, (36.4m and 7.1 points respectively therefore cannot be recommended for use at the individual level. Due to ordinal nature of the SCS, measurement error values were unable to be calculated for this measure and Kappa statistic was 0.314 ICC was reported as 0.728 and 0.704 for the EQ-5D-index and VAS respectively, therefore more data is required to be collected on the EQ-5D-5L in pwLLAs, before its reliability can be confirmed in this population.

It is recognised that the MDC represents half the LoA calculated for the same population (Norman et al. 2003) and therefore the recommendations for measurement error are similar to those put forward for the agreement parameters above.

The results obtained, identified that only the SIGAM and TUG demonstrated excellent reliability over repeated measures for *both* consistency and agreement with ICC (2,1) values > 0.9 and no issues identified from the Bland Altman plots with sufficiently narrow LoA (SIGAM: 1.2 grade and TUG: 3.4s). A MDC value had not been presented for the SIGAM before, and with excellent consistency and agreement there is confidence that this MDC_{95} of one grade (out of 11 possible grades) can be applied to the individual. Therefore a change of one grade on the SIGAM or more is sufficient to demonstrate a "real", change (i.e. not due to inherent variability) in the functional ability of a pwLLA. This result is presented to a confidence level of 95% which means that if the same population is sampled on numerous occasions this MDC value (of one) would be found in approximately 95% of the cases.

On the basis of the excellent consistency and agreement results for the TUG, the MDC_{95} of 1.8s can also be recommended to clinicians as the minimum value required to demonstrate a real change (in the balance and mobility) in an *individual* pwLLA. A MDC_{90} value of 3.6s (the equivalent MDC_{95} value would be higher) had previously been presented for the TUG by Resnik et al in 2011, but with no comment on the acceptability of the LoA in the results or discussion it was not clear whether this MDC value was acceptable to be applied to changes seen in a individual patient. However, the authors did recommend that clinicians look for a change of at least 3.6s between repeated results in order to be sure that a "real", i.e. beyond measurement error, change had occurred in their patient. Therefore inferring that the outcome measure used can detect a meaningful difference at the individual level (Resnik and Borgia 2011). If no comment is made on limits of agreement it is better to assume that a larger measurement error is required than the one calculated from the group; and that changes greater than this must be detected for the clinician to be sure their patient has made a true change i.e. beyond random error (de Vet et al. 2011).

Excellent consistency was demonstrated for the LCI-5 and 2MWT with ICC values also over 0.9. However, the evidence for agreement was less convincing for both. No MDC values have been previously presented for the LCI-5 either, but while the LCI-5 demonstrated excellent consistency (ICC>0.9) the wide LoA of 7.1 (range -3.3 - 3.8 points) questioned the appropriateness of this outcome measure to assess change at the individual level. The 2MWT also demonstrated excellent consistency (ICC>0.9) but the appropriateness of this outcome measure to assess change at the individual level was also questioned due to the wide LoA (36.4m). Therefore, the MDC_{95} of 4 on the LCI-5 and 18.9m for the 2MWT, cannot be recommended for use at the individual level, but may be used in group level analysis on the basis of the results from this study.

Good consistency (ICC >0.7 and <0.9) was reported for both the EQ-5D index scores and VAS within repeated measures. However, agreement was difficult to comment on as this was the first presentation of reliability for the EQ-5D-5L in a population of pwLLA and it was not clear whether the agreement limits reported were representative of reasonable variability in the clinical performance of this population. A change score of at least 0.258 (MDC_{95}) on the EQ-5D index score and at least 15.5 on the EQ-5D VAS was calculated to be change required to be confident that a real change, has occurred in the global health of a pwLLA. However, with agreement difficult to comment on as discussed above and despite the good consistency seen for both the index score and the VAS, the collection of more data in clinical practice or research is recommended so that the values of the agreement parameters for the EQ-5D-5L can be put in to context with this population. As these are the first measurement error values presented for the EQ-5D-5L in pwLLA, it is hoped that they will provide reference values for researchers planning future studies. Confirmation of the clinical applicability of these values could be gathered by comparing them to values reported in other studies on this population.

Finally, with only fair consistency recorded for the SCS and 50% agreement between the two visits, neither the consistency nor agreement of this outcome measure can be confirmed on the basis of these results.

A clinician or researcher should not use an outcome measure to evaluate the effectiveness of an intervention without knowledge of its reliability (including their measurement error values). This is because they cannot know if the patient's performance (as measured by that outcome measure) has truly changed, i.e. is "real" and beyond measurement error. The new data provided from the repeatability study on the reliability and measurement error values for these outcome measures will provide useful for clinicians and researchers in their work going forward in the context of prosthetic rehabilitation.

7.3.3.2 Responsiveness

In study IV (the longitudinal cohort study reported in chapter 6), thirty unilateral pwLLAs (adults) were followed during the early phase of prosthetic rehabilitation, when changes in the physical ability and function of the participants were expected. Study visits occurred: approximately one week after they were fitted with their first prosthetic limb; two weeks later and; to coincide with their MDT clinic appointment, six weeks after hospital discharge. Repeated measures ANOVA and effect size were calculated for the SIGAM, LCI-5, TUG, 2MWT, SCS and the EQ-5D-5L. In addition, an anchor-based method of calculating external responsiveness was employed. The anchor used was a change questionnaire focused on the participants' perceived ability of undertaking physical tasks and the changes in that ability between milestones, i.e. delivery of prosthesis, discharge from hospital and at the time of their 6-week post-discharge clinic appointment. Minimal clinically important differences were calculated for each outcome measure as an index of responsiveness for each time interval. The key findings from this study are presented in Table 7.2 below.

Table 7.2 Key findings from Study IV

Research Questions (study/chapter)	**Key Findings**
Question 6: What is the responsiveness of Physical Function outcome measures	All outcome measures, except the SCS and EQ-5D-5L, were responsive to changes across the whole study period, as evidenced from the ANOVA and effect size results.

regularly used with lower limb amputees when assessing change during the rehabilitation period immediately following limb fitting? (Longitudinal cohort study - study IV / chapter 6)	In addition to recording significant differences in both time intervals, the LCI-5, and 2MWT recorded larger effect sizes in the in-patient period when compared to the post-hospital discharge period, 0.66 vs 0.53 and 0.69 vs 0.47 respectively. The effect sizes were similar in both the in-patient and post-hospital discharge periods for the TUG. Validity testing for the ACQ was confirmed (by achieving more than 50% of the criteria) **for only one** outcome measure, the SIGAM in TI1. The lack of confirmation of the validity for the ACQ also weakened the support for the MCID values. Therefore, the values that have been presented should be used with caution and may have limited clinical use. MCID values were established for all outcome measures during all time intervals. However, results of the hypothesis testing did not support the accuracy of any of the outcome measures, when used with the ACQ as a measure of patient perception of change

Using the “distribution-based” approach for estimating responsiveness, (i.e. ANOVA) all outcome measures were considered responsive to change across the whole study period (TI3, which = TI1 + TI2). Post-hoc tests showed statistically significant differences between the first and last study visits for all outcome measures except the SCS and EQ-5D-index scores

When examining the magnitude of the changes detected in the different time intervals which were statistically significant, the LCI-5 and 2MWT recorded larger effect sizes in the immediate post limb-fitting period (TI1) when compared to the post-hospital discharge period (TI2). The TUG recorded virtually identical effect sizes in both. The effect sizes in TI1 were medium for all (0.66 - LCI-5, 0.53 – TUG and 0.69 - 2MWT), and in TI2 were either medium (0.53 – LCI-5 and 0.54 - TUG) or small (0.47 – 2MWT). However, these results alone would earn the study a rating of “fair” for the responsiveness measurement property within the COSMIN methodological quality checklist for only presenting effect size or results of inferential statistical tests.

No MCID values have been previously presented for pwLLA for any of the outcome measures in study IV. According to the COSMIN guidelines, any values for responsiveness derived for the outcome measures in study IV must be considered in the context of a comparator instrument. In this study the patient reported ACQ or clinical 'anchor' was considered a "comparator instrument" which recorded the pwLLAs' perception of change with respect to their physical ability to undertake everyday tasks. For a comparator instrument to be considered acceptable and achieve an excellent rating for methodology, according to the COSMIN checklist, both the instrument and its measurement properties should be adequately described. While the ACQ was adequately described within the methodology for study IV, its psychometric properties were not confirmed. The results, detailed in chapter 6 (section 6.4.7) did not confirm validity of the ACQ when used with any of the outcome measures being studied. Therefore it is likely that a rating of poor may apply, i.e. "some information on measurement properties of the comparator instrument(s) in any study population", but not a good rating: because "adequate" measurement properties of the comparator instrument had not been confirmed but not sure if these apply to the study population" (see Appendix 4.3 for COSMIN checklist). Further discussion of the use of change questionnaires in the analysis of responsiveness will continue in section 7.4.1.2 below.

The accuracy of the outcome measures in correctly identifying individuals who had and who had not undergone an important change based on the responses of the ACQ (much better vs better, same, worse and much worse) was no better than moderate and this was only seen in TI2 The moderately accurate findings were presented in TI2 for the SIGAM, TUG, 2MWT and EQ-5D-VAS are likely to be compromised due to the very small numbers in the "other" group (n=3) compared to the "much better group" (n=20) in the ROC curve analysis. The remainder of the ROC curve analysis results demonstrated only low accuracy or chance results. For example the accuracy results across the whole study period (TI3) was slightly better than chance, i.e. low accuracy, except when used with the 2MWT and the SCS where it was seen as a chance result. Therefore the MCID results presented for SIGAM (5 grades), LCI-5 (14 points), TUG (5.5s) and EQ-5D index (0.168) and VAS (7.5 points) for the whole study period (TI3) can only be confirmed and recommended with caution.

7.4 Recommendations for future studies

Several challenges (or limitations), have been identified throughout this thesis, and are tabulated later in the chapter (see Table 7.3 in section 7.5) together with future recommendations. However, some of these are worthy of further scrutiny and will be examined in more detail in the following sections.

7.4.1 Choice of statistical analysis

Inappropriate statistical analyses and incomplete reporting of results for the psychometric properties of an outcome measure may make it difficult to interpret the evidence presented. Consequently, incorrect interpretation of the evidence may result in an inappropriate outcome measure being used, which in turn will have implications on the reported effect of an intervention. For example, the effects of an intervention may be under or over estimated when ordinal scores are used inappropriately (Stucki et al. 1996). Ordinal scales, such as Likert scales, of which the SCS is a numerical example, will only record an order of magnitude. It is not known how much better one level is than the one below or above, as the difference between the levels may not be considered equal by everyone using the scale. Correct use of ordinal data is also highlighted in Rethinking Rehabilitation, Theory and Practice where clinicians are encouraged not to use ordinal scales to evaluate the effectiveness of an intervention or therapy unless the scale had undergone a Rasch analysis (McPherson et al. 2015). The following sections will address issues with the analysis and reporting choices for reliability and responsiveness, which are the main focus of the two interventional studies (III and IV).

7.4.1.1 Reliability: consistency vs agreement

It has been recognised that both consistency and agreement parameters are presented, as neither alone provides sufficient information on the reliability of an outcome measure to recommend its utility in rehabilitation studies (Rankin and Stokes 1998). Consistency parameters are highly dependent on the heterogeneity of the characteristics of the performance of the study sample, whereas agreement parameters are more dependent on the characteristic of the measurement instrument and are (directly) related to the measurement error (de Vet et al. 2006).

However, there appeared to be a tendency in the literature studied for the systematic review in this thesis, for consistency parameters to be presented without commenting on agreement (systematic review in chapter 4). High ICC values were presented without reference to agreement, and authors declared that the outcome measure in question demonstrated excellent or good reliability e.g. Wong et al 2013 for the BBS, Brooks et al (2002) for the 2MWT and Resnik et al 2011 for 2MWT, TUG and others. Following the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (Kottner et al. 2011), the recommendations and conclusions from the results in chapter 5 made comment on both consistency and agreement parameters for all the outcome measures studied. Therefore, allowing the clinician or researcher to make a fully informed choice from all the evidence for the reliability of these outcome measures.

7.4.1.2 Responsiveness: distribution-vs anchor-based methods

Throughout this thesis, responsiveness has been considered a separate measurement property, distinct from validity. The definition of this separation, proposed by Guyatt et al 1987, states that validity refers to a single score and responsiveness considers a change score based on two measurements. This definition underpinned the study design and methodology in study IV (the longitudinal cohort study) and is supported by others (Mokkink et al. 2010, Beaton 2001, Husted et al. 2000). In 2000 Husted et al (2000) spoke of a distinction between internal and external responsiveness; with "internal" responsiveness characterised as the ability of a measure to change over a particular pre-specified time frame and "external" responsiveness as the extent to which changes in a measure over a specified time frame related to corresponding changes in a reference measure of health status. What Husted et al (2000) referred to as "internal responsiveness" of an outcome measure, is presented using statistical methods of analysis, e.g. ES, SRM, paired t-test, GRI. These methods are also known as "distribution-based" methods of calculating responsiveness. It is possible to show that an outcome measure can detect change by demonstrating a statistically significant change in group mean change scores; and that the change did not occur by chance or was a measurement error. This is why the COSMIN checklist requires the presence of an appropriate comparator measure (i.e. proof of external responsiveness) for an excellent rating in the responsiveness measurement property category.

Despite the continued debate about which of the above methods should be used to measure responsiveness, possibly fuelled by the continuing division of opinion on its definition and separation from validity, it is now recommended by many (Haley and Fragala-Pinkham 2006, Copay et al. 2007, Revicki et al. 2008) that results from distribution-based methods are used to support anchor- or criterion-based results. This is because distribution-based indices provide no direct information about level of clinical importance, they are simply a way of expressing an observed change in a standardised metric (Hays et al. 2005). However there continues to be a lack of evidence on the ability of outcome measures to detect these "clinically" important changes in many populations. For example, Scrivener et al (2013) published a systematic review on the responsiveness of measurement tools in the stroke population (Scrivener et al. 2013). This was after finding in previous reviews (Salter et al. 2005, Tyson 2009) that responsiveness had rarely been investigated. Scrivener et al (2013) concluded that the lower limb physical performance measures, included in the review, did demonstrate "generally large" responsiveness (using distribution-based methods); however no comment was made on whether clinically important changes were detected.

It is also recognised that changes detected by an outcome measure must be relevant not only to the patient but also the context before it can be deemed responsive (for that population and in that context) (Haley and Fragala-Pinkham 2006, Copay et al. 2007, Terwee et al. 2003). Calculating the MCID values using the anchor-based method and ROC analysis is one approach to investigating how much change is enough (or clinically important), and was the approach chosen in this thesis. However, it may not only be concerns about which methods to use (Angst 2011) that are creating this continuing gap in the evidence for responsiveness. Difficulties of implementing anchor-based methods in the area of healthcare and rehabilitation have been noted. Validity of the anchor used to calculate MCID values must be considered to legitimise the values that are calculated using it (Gatchel and Mayer 2010). However, as was seen in study IV, validation of the anchor used was challenging. The number of points on the scale and the specificity of the anchor question were identified as possible causes of the poor validity results. While some authors have noted similar problems (Rushton et al. 2015), others have found good validity when using the Global Rating of Change

(GRC) (Kamper 2009). Examination of different examples of GRC scales used discovered that there was a wide variation in the type of question asked, the number of points on the scale, and also the labels that were assigned to the scale points (Kamper 2009). Variations in study design e.g. the time interval over which the change is being assessed and the method in which it is collected, may also affect the validity of the scale being used as an anchor, because of recall bias. Kamper et al (2009) provide some advice from a review of the literature on both the wording of the question used and the size of the scale. Firstly, the question asked must direct the patient / participant / clinician towards the construct being measured by the outcome measure under investigation. Secondly, a specific time interval must be given so that the respondent can "anchor" their response to a time point from which he / she can compare the current status. And finally, if both improving and deteriorating changes are to be collected then the scale should be balanced around zero. A seven level scale was found to offer the best sense of balance between acceptability by the respondent (i.e. not too long), adequate discrimination between levels of change and test re-test reliability (Kamper 2009).

Both global and specific change questionnaires have been used when trying to establish MCID values using a patient anchor. The GRC, discussed briefly in the previous section is an example of a global health change questionnaire used in responsiveness studies. For example, the GRC was used trying to estimate how well the L Test identified LLAs who had or had not undergone a clinically important change in their ability to get up and walk with their prosthesis (Rushton et al. 2015). The AUC was 0.67 (representing low accuracy of the L Test when using the GRC), therefore the authors concluded that the MCID value that was presented should be interpreted with caution. The mRS (modified Rankin Score) is an example of a specific health outcome which measures the level of disability in people after a stroke and other neurological conditions. The mRS was used in the study by Tilson et al in 2010 estimating the MCID value for comfortable gait speed (CGS) (Tilson et al. 2010). The AUC was 0.69 and therefore there only low accuracy was also recorded for CGS. The authors noted that the failure to identify a clinically important improvement in CGS when using the mRS was probably because the mRS was not a gait-specific measure of disability. However, it is expected that either a global or specific health questionnaire could be used to good effect as an appropriate anchor

questionnaire as long as the specific advice provided by Kamper et al (2009) (outlined earlier in the previous paragraph) is followed.

7.4.2 Patient reported vs objectively measured outcome measures

How an outcome measure is administrated, i.e. whether self-reported or observed will have an effect on the results. Psychosocial factors such as low self-efficacy and self-confidence may influence the results of self-reported outcome measures or PROMs. Collecting both PROMs and observed outcome measures at the same time may shed light on any such influences. While, an analysis of such correlations between the changes recorded in the observed measures and those in the PROMs was not undertaken in this thesis, links between the two have been explored previously, especially in the area of chronic pain and mental health. For example, Wittink et al (2003) found that the self-report measure SF-36 physical functioning (PF) scale and a treadmill test (a performance-based assessment) provided information about distinct, although related, domains of physical functioning in patients with chronic low back pain (Wittink et al. 2003). The strongest relationship was found between disability status and mental health (as measured by the SF-36). They also determined poor mental health was associated with lower SF-36 (PF) scores but the treadmill data was not correspondingly low, thus demonstrating a discrepancy between the self-reported and performance data people with poor mental health. A recognised factor in this relationship is depression. Given that the prevalence of depression (and anxiety) is relatively high in pwLLAs during the first two year post-amputation (Horgan and MacLachlan 2004), there is the possibility that the same relationship may exist between mental and physical health in pwLLAs. In 2010 Parker et al, published a study on the relationship between capacity and performance measures with specific reference to ambulation of people with pwLLAs in a community setting (Parker et al. 2010). In addition to establishing a high correlation between the 2MWT and a step activity monitor, they also saw a negative relationship between depressive symptoms and performance. However, depression was only related to decreased performance as measured by the Trinity Amputation and Prosthesis Experience Scale (TAPES) Activity Restriction subscales (a self-report measure (Gallagher and MacLachlan 2004)) and not the observed measures of performance.

Data from both PROMs and observed performance should be taken into account when considering the impact of a prosthetic component. High correlation between

perceived and actual performance would lend credibility to any improvements observed or perceived. The review by Kannenburg et al (2014) looked at the benefits of micro-processor-controlled knees joints in community walkers. However, improvements in perceived function and mobility as assessed by self-reported outcome measures for use of a MPCK lagged behind the performance-based improvements with only 22 percent of the self-reported outcome measures related to function, mobility, and satisfaction showing significant improvements as a result of MPK use. The findings were somewhat surprising to the authors who expected the observed improvements in function to be reflected more clearly in the pwLLAs' perceptions. However, it appeared that lower performing pwLLAs (as classified by the Medicare K-levels), did not subjectively report the improvements that were seen in objective performance by MPK use. Both self-report and performance based outcome measures were also used in a study to determine functional differences between four categories of prosthetic feet. While significant differences were found in some groups using the Amputee Mobility Predictor with a prosthesis (AMPPRO), a performance measure, neither of the self-report measures (PEQ-13 and LCI-5) used in the study were able to detect differences between prosthetic feet (Gailey et al. 2002). However, this is unlikely to have been due to lower performing pwLLAs perceiving less improvement, as found in the review by Kannenburg et al (2014), but is probably due to the high scores recorded for the PEQ-13 and LCI-5 throughout the study. The authors noted the "ceiling" effect of both these measures (Gailey et al. 2002).

The evaluative evidence presented in these, and other similar studies, is equivocal about which prosthetic component may be "better" for different groups of pwLLAs. Whether a self-reported or objectively measured outcome measure is being used, the question must be asked: How robust are the results from these studies if there is no strong evidence for the "external" responsiveness for any of the outcome measures being used? A second question may also be asked: Are the "correct" (i.e. appropriate) constructs being measured in order to show differences across such a heterogenic population? Whatever type of outcome measure is being used, the constructs being investigated are dictated by the assessor. Maybe the person being assessed should choose the construct of interest, i.e. the performance, task, attitude that he/she would like to see a change in? This could be done by patient-centred outcome measures (PCOMs), which can be regarded as PROMs because they are

reported by the patient. However, in PCOMs the constructs being measured are identified BY the patient and not restricted. This makes them distinct from other PROMs, where the constructs being measured are decided FOR the patient and are restricted to what the outcome measure was developed to measure. Therefore PCOMs can address any aspects covered in the ICF including any psychosocial aspects, if considered important to the amputee. This type of outcome measure is being promoted in many areas of rehabilitation where the patient may have complex needs, including the area of pain management (O'Connell et al. 2015). This approach has been investigated with pwLLAs, and the psychometric properties of an outcome measure of this type, the Goal Attainment Scale (GAS), has been reported by Rushton et al (2002). However, the GAS does not seem to be regularly reported in the research. Goal Attainment Scales may provide an alternative way to measuring the effectiveness of a treatment or intervention that is specific to the individual patient and thereby avoiding the use of multiple outcome measures. There are practical guides to help implement GAS within clinical teams (Steenbeek et al. 2008, Turner-Stokes 2009). Some guidance has been published on the mathematical analysis when using GAS (Tennant 2007), and a literature-based update on the use of GAS in rehabilitation was published in 2013 (Krasny-Pacini et al. 2013). This outlined the utility of GAS across many areas of rehabilitation for multiple roles. However, the authors noted that despite these many reviews there is no robust evidence for the psychometric qualities in clinical practice or research (Krasny-Pacini et al. 2013). This was due, in part, to the lack of information on the methodological quality of the studies (i.e. no standardised checklist such as COSMIN was used) and variations in methodologies which could impact on the validity of the GAS in different situations and also the interpretation of the results.

7.5 Limitations and recommendations for future work

Table 7.3 Main limitations identified throughout thesis

Limitations	**Recommendations**
The sample size in the survey (study I – chapter 3) was limited in some groups (prosthetists, occupational therapist and pwLLAs) and it was conducted just over three years prior to thesis submission.	A repeat survey, with a robust distribution plan would be recommended to reach a wider audience to confirm and/or update the results.
Studies investigating outcome measures used exclusively for measuring quality of life were excluded from the systematic review of the literature. In addition, the degree to which psychosocial factors impacted on the pwLLA following his/her amputation was not considered during the conduct of the interventional studies.	It is now recognised that these factors have the potential to influence results obtained from both observed and self-reported outcome measures. Therefore psychosocial factors should be recorded in any future studies to establish any confounding influences on the results.
While the COSMIN checklist was developed for use with HR-PRO questionnaires, it was used in study II (the systematic review - chapter 4) with observed measures of performance as well as multi-item questionnaires.	The theoretical framework underpinning the development of the COSMIN checklist was applicable for the assessment of the psychometric properties of observed measures of performance as well as HR-PRO questionnaires.
This was a single researcher PhD and therefore inter-rater reliability was not addressed in either study III (chapter 5) or study IV (chapter 6)	The use of two raters, randomly assigned to participants and visits will provide the opportunity for both inter and intra-rater reliability analysis.
The sample sizes were such that in both study III and study IV, the methodological quality of the study would be considered “poor” and “fair”, respectively, if assessed using the COSMIN quality criteria, and the strength of evidence would be deemed indeterminate.	Collaborative multi-centre studies will increase the recruitment pool and subsequent sample size.

The outcome measures investigated (either patient-reported or observed) were measures where constructs being measured were restricted to a single construct and dictated by what the outcome measure was developed to measure.	The use of patient-centred outcome measures, in addition to uni-dimensional measures, will capture changes in constructs identified by the patient and multi-dimensional effects could be measured.
Recruitment of both TF and TT amputees was a potential source of systematic variation in both study III and IV.	As above, the use of collaborative multi-centre studies would increase the potential recruitment pool and a larger sample size would allow for greater numbers in any sub-group analyses.
The ACQ used in study IV was a five-level response questionnaire which was considered too limited to distinguish what is an important change. The limited (and uneven) numbers in each group for the ROC curve analysis reduced the ability to detect the accuracy of the outcome measures.	The use of a questionnaire with more than 7 levels may detect smaller incremental changes. This coupled with a larger sample will produce larger, and more even numbers in each group for the ROC curve analysis.
The question associated with the ACQ was not specific to each outcome measure. The validity of the ACQ may have been improved if specific questions relating to the different constructs of each outcome measure had been included.	Specific questions for each construct being measured should be devised when using an anchor-based method to estimate the responsiveness of an outcome measure.

7.6 Conclusions

There is still no consensus on which outcome measures to use with pwLLAs in prosthetic rehabilitation. The most common outcome measures used in prosthetic rehabilitation in the UK measure constructs in the physical domain (chapter 3).

The need to use outcome measures that have robust psychometric properties to evaluate the effectiveness of interventions such as therapy and gait training programmes as well as prosthetic componentry is well recognised. However, very few outcome measures that measure physical function, being used with pwLLAs in prosthetic rehabilitation, have high quality evidence, especially with regard to knowledge of their measurement error and their responsiveness (chapter 4).

The results from both the repeatability study (chapter 5) and the longitudinal cohort study (chapter 6) have added to the evidence base for the psychometric properties for the most commonly used outcome measures measuring physical function in early prosthetic rehabilitation. Further work is still required to confirm or refute these findings, largely due to the small sample sizes reducing the strength of the evidence presented.

Difficulties were highlighted when using a patient anchor to establish external responsiveness of the outcome measures. It is recommended that appropriately worded anchor questions specific to the construct being measured are constructed for any future work using change questionnaires as the anchor. The provision of at least three levels of responses for improvement and deterioration should also be considered.

Recommendations have been presented for both clinicians and researchers regarding the use of outcome measures, with individuals and for groups in prosthetic rehabilitation and research. However, some outstanding questions still remain surrounding the evidence for the use of these outcome measures to evaluate the effectiveness of an intervention. Fundamental to these recommendations is the use of standardised guidelines and checklists when developing outcome measures, as well as when designing and reporting on studies that investigate the psychometric

properties of an established outcome measure in a new population. Thus the choice of any outcome measure will be underpinned by strong evidence

Outstanding questions also remain about the types of outcome measures that should be used and whether patient-report, patient centred or observed measures can capture clinically important changes when evaluating the effectiveness of any intervention. Whichever individual outcome measure, or combination of measures is used, the construct(s) being measured must be one(s) that is (are) of concern to the "investigator". It has been argued that the "investigator" may be a clinician, researcher or the patient.

Fundamental to the choice of outcome measures, whoever the investigator is, is that the psychometric properties of the measure have been established for its use with the particular population of patients, and in the particular context it is about to be used.

The following quote, often (wrongly) attributed to Einstein, may be useful to remember when planning future research in this field:

"Not everything that counts can be counted and
not everything that can be counted counts"
(Cameron 1963)

References

ABRAMS, D., DAVIDSON, M., HARRICK, J., HARCOURT, P., ZYLINSKI, M. and CLANCY, J., 2006. Monitoring the change: current trends in outcome measure usage in physiotherapy. *Manual Therapy.* vol. 11, no. 1, pp. 46-53.

AHMAD, N., THOMAS, G.N., GILL, P., CHAN, C. and TORELLA, F., 2014. Lower limb amputation in England: prevalence, regional variation and relationship with revascularisation, deprivation and risk factors. A retrospective review of hospital data. *Journal of the Royal Society of Medicine.* vol. 107, no. 12, pp. 483-489.

ALVIAR, M.J., OLVER, J., BRAND, C., TROPEA, J., HALE, T., PIRPIRIS, M. and KHAN, F., 2011. Do patient-reported outcome measures in hip and knee arthroplasty rehabilitation have robust measurement attributes? A systematic review. *Journal of Rehabilitation Medicine.* vol. 43, no. 7, pp. 572-583.

AMMANN-REIFFER, C., BASTIAENEN, C.H.G., DE BIE, R. and VAN HEDEL, H.,J, 2014. Measurement Properties of Gait-Related Outcomes in Youth With Neuromuscular Diagnoses: A Systematic Review. *Physical Therapy.* vol. 94, no. 8, pp. 1067-1082.

ANDRESEN, E.M., 2000. Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine and Rehabilitation.* vol. 81, no. 12, pp. 15-20.

ANGST, F., 2011. Correspondence: The new COSMIN guidelines confront traditional concepts of responsiveness. *BioMedCentral Medical Research Methodology.* vol. 11, pp. 152-157.

ANGST, F., AESCHLIMANN, A. and STUCKI, G., 2001. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis & Rheumatism: Arthritis Care & Research.* vol. 45, no. 4, pp. 384-391.

BANDURA, A., 2010. Self-Efficacy. In: *The Corsini Encyclopedia of Psychology.* John Wiley & Sons, Inc., DOI 10.1002/9780470479216.corpsy0836.

BARTELS, B., DE GROOT, J.F. and TERWEE, C.B., 2013. The Six-Minute Walk Test in Chronic Pediatric Conditions: A Systematic Review of Measurement Properties. *Physical Therapy.* vol. 93, no. 4, pp. 529-541.

BAUTISTA-MOLANO, W., NAVARRO-COMPAN, V., LANDEWE, R.,B.M., BOERS, M., KIRKHAM, J.J. and VAN, D.H., 2014. How well are the ASAS/OMERACT core outcome sets for ankylosing spondylitis implemented in randomized clinical trials? A systematic literature review. *Clinical Rheumatology.* vol. 33, no. 9, pp. 1313-1322.

BEATON, D., 2001. A taxonomy for responsiveness. *Journal of Clinical Epidemiology.* vol. 54, no. 12, pp. 1204-1217.

BECKERMAN, H., ROEBROECK, M.E., LANKHORST, G.J., BECHER, J.G., BEZEMER, P.D. and VERBEEK, A.L.M., 2001. Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research.* vol. 10, no. 7, pp. 571-578.

BHERER, L.E.,K. and LIU-AMBROSE, T., 2013. A Review of the Effects of Physical Activity and Exercise on Cognitive and Brain Functions in Older Adults. *Journal of Aging Research.* vol. 2013, pp. 1-8.

BLACK, N., 2013. Patient reported outcome measures could help transform healthcare. *British Medical Journal.* vol. 346, no. 1, pp. f167.

BLAND, J. and ALTMAN, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet.* vol. 327, no. 8476, pp. 307-310.

BOERS, M., KIRWAN, J.R., WELLS, G., BEATON, D., GOSSEC, L., D'AGOSTINO, M., CONAGHAN, P.G., BINGHAM, C.O., BROOKS, P., LANDEWE, R., MARCH, L., SIMON, L.S., SINGH, J.A., STRAND, V. and TUGWELL, P., 2014. Developing Core Outcome Measurement Sets for Clinical Trials: OMERACT Filter 2.0. *Journal of Clinical Epidemiology.* vol. 67, no. 7, pp. 745-753.

BOHANNON, R.W., 2006. Reference Values for the Timed Up and Go Test. *Journal of Geriatric Physical Therapy.* vol. 29, no. 2, pp. 64-68.

BOHANNON, R.W. and GLENNEY, S.S., 2014. Minimal clinically important difference for change in comfortable gait speed of adults with pathology: a systematic review. *Journal of Evaluation in Clinical Practice.* vol. 20, no. 4, pp. 295-300.

BOHANNON, R.W., WANG, Y. and GERSHON, R.C., 2015. Two-Minute Walk Test Performance by Adults 18 to 85 Years: Normative Values, Reliability, and Responsiveness. *Archives of Physical Medicine and Rehabilitation.* vol. 96, no. 3, pp. 472-477.

BRAZIER, J., DEVERILL, M., GREEN, C., HARPER, R. and BOOTH, A., 1999. A review of the use of health status measures in economic evaluation. *Health Technology Assessment.* vol. 3, no. 9, pp. 1-164.

BROOKS, D., HUNTER, J.P., PARSONS, J., LIVSEY, E., QUIRT, J. and DEVLIN, M., 2002. Reliability of the two-minute walk test in individuals with transtibial amputation. *Archives of Physical Medicine and Rehabilitation.* vol. 83, no. 11, pp. 1562-1565.

BROOKS, D., PARSONS, J., HUNTER, J.P., DEVLIN, M. and WALKER, J., 2001. The 2-minute walk test as a measure of functional improvement in persons with lower limb amputation. *Archives of Physical Medicine and Rehabilitation.* vol. 82, no. 10, pp. 1478-1483.

BUTLAND, R.J., PANG, J., GROSS, E.R., WOODCOCK, A.A. and GEDDES, D.M., 1982. Two-, six-, and 12-minute walking tests in respiratory disease. *British Medical Journal (Clinical Research Ed.).* vol. 284, no. 6329, pp. 1607-1608.

BUTLER, K., BOWEN, C., HUGHES, A.M., TORAH, R., AYALA, I., TUDOR, J. and METCALF, C.D., 2014. A systematic review of the key factors affecting tissue viability and rehabilitation outcomes of the residual limb in lower extremity traumatic amputees. *Journal of Tissue Viability.* vol. 23, no. 3, pp. 81-93.

CALLAGHAN, B.G., SOCKALINGAM, S., TREWEEK, S.P. and CONDIE, M.E., 2002. A post-discharge functional outcome measure for lower limb amputees: test-retest reliability with trans-tibial amputees. *Prosthetics and Orthotics International.* vol. 26, no. 2, pp. 113-119.

CAMERON, W.,B., 1963. *Informal Sociology: A Casual Introduction to Sociological Thinking.* New York: Random House.

CANO, S. and HOBART, J., 2011. The problem with health measurement. *Patient Preference and Adherence.* vol. 5, pp. 279-290.

CASTIEN, R.F., BLANKENSTEIN, A.H., WINDT, D.A.W.M.v.d. and DEKKER, J., 2012. Minimal clinically important change on the Headache Impact Test-6 questionnaire in patients with chronic tension-type headache. *Cephalalgia: An International Journal of Headache.* vol. 32, no. 9, pp. 710-714.

CHAN, L., HEINEMANN, A.W. and ROBERTS, J., 2014. Elevating the quality of disability and rehabilitation research: Mandatory use of the reporting guidelines. *American Journal of Occupational Therapy.* vol. 68, no. 2, pp. 127-129.

CICCHETTI, D.V., 1994. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment.* vol. 6, no. 4, pp. 284-290.

CIEZA, A., EWERT, T., BERDIRHAN USTUN, T., CHATTERJI, S., KOSTANJSEK, N. and STUCKI, G., 2004. Development of ICF Core Sets for patients with chronic conditions. *Journal of Rehabilitation Medicine.* vol. 36, pp. 9-11.

COHEN, J., 1988. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

COLE, M.J., CUMMING, J., GOLLAND, N., HAYES, S., OSTLER, C., SCOPES, J.and TISDALE, L., 2014. *BACPAR TOOLBOX OF OUTCOME MEASURES Version 2.* Professional Network Report ed. British Association of Chartered Physiotherapists in Amputee Rehabilitation.

CONDIE, E., SCOTT, H. and TREWEEK, S., 2006. Lower limb prosthetic outcome measures: A review of the literature 1995 to 2005. *Journal of Prosthetics and Orthotics.* vol. 18, no. 6, pp. 13-45.

CONDIE, M.E., MCFADYEN, A.K., TREWEEK, S. and WHITEHEAD, L., 2011. The trans-femoral fitting predictor: a functional measure to predict prosthetic fitting in transfemoral amputees--validity and reliability. *Archives of Physical Medicine and Rehabilitation.* vol. 92, no. 8, pp. 1293-1297.

COPAY, A.G., SUBACH, B.R., GLASSMAN, S.D., POLLY, D.W. and SCHULER, T.C., 2007. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal.* 09, vol. 7, no. 5, pp. 541-546.

CORR, S. and SIDDONS, L., 2005. An introduction to the selection of outcome measures. *British Journal of Occupational Therapy.* vol. 68, no. 5, pp. 202-206.

DAVIE-SMITH, F., PAUL, L., NICHOLLS, N., STUART, W.P. and KENNON, B., 2016. The impact of gender, level of amputation and diabetes on prosthetic fit rates following major lower extremity amputation*. Prosthetics and Orthotics International* pp. 1-7 Available from: http://poi.sagepub.com/lookup/doi/10.1177/0309364616628341 DOI 10.1177/0309364616628341.

DE LAAT, F.,A., ROMMERS, G.M., GEERTZEN, J.H. and ROORDA, L.D., 2012. Construct validity and test-retest reliability of the walking questionnaire in people with a lower limb amputation. *Archives of Physical Medicine and Rehabilitation.* vol. 93, no. 6, pp. 983-989.

DE LAAT, F.,A., ROMMERS, G.M., GEERTZEN, J.H. and ROORDA, L.D., 2011. Construct validity and test-retest reliability of the questionnaire rising and sitting down in lower-limb amputees. *Archives of Physical Medicine and Rehabilitation.* vol. 92, no. 8, pp. 1305-1310.

DE LAAT, F.,A., ROMMERS, G.M., GEERTZEN, J.H. and ROORDA, L.D., 2010. Construct validity and test-retest reliability of the climbing stairs questionnaire in lower-limb amputees. *Archives of Physical Medicine and Rehabilitation.* vol. 91, no. 9, pp. 1396-1401.

DE VET, H.,C., TERWEE, C.B., OSTELO, R.W., BECKERMAN, H., KNOL, D.L. and BOUTER, L.M., 2006. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes.* vol. 4, pp. 54-54.

DE VET, H.,C.W., TERWEE, C.B., KNOL, D.L. and BOUTER, L.M., 2006. When to use agreement versus reliability measures. *Journal of Clinical Epidemiology.* vol. 59, no. 10, pp. 1033-1039.

DE VET, H.,C.W., TERWEE, C.B., MOKKINK, L.B. and KNOL, D.L., 2011. *Measurement in Medicine: A practical guide.* 1st edition ed. United Kingdom: Cambridge University Press.

DEATHE, A.B., WOLFE, D.L., DEVLIN, M., HEBERT, J.S., MILLER, W.C. and PALLAVESHI, L., 2009. Selection of outcome measures in lower extremity amputation rehabilitation: ICF activities. *Disability and Rehabilitation.* vol. 31, no. 18, pp. 1455-1473.

DEATHE, A.B. and MILLER, W.C., 2005. The L test of functional mobility: measurement properties of a modified version of the timed "up & go" test designed for people with lower-limb amputations. *Physical Therapy.* vol. 85, no. 7, pp. 626-635.

DEFENCE, M.o., 2013. *Quarterly Afghanistan and Iraq Amputation Statistics.* Defence Statistics (Health) no. July 2015, pp. 1-10.

DEMET, K., MARTINET, N., GUILLEMIN, F., PAYSANT, J. and ANDRÉ, J., 2003. Health related quality of life and related factors in 539 persons with amputation of upper and lower limb. *Disability and Rehabilitation.* vol. 25, no. 9, pp. 480-486.

DES JARLAIS, D.C., LYLES, C. and CREPAZ, N., 2004. Improving the Reporting Quality of Nonrandomized Evaluations of Behavioral and Public Health Interventions: The TREND Statement. *American Journal of Public Health.* vol. 94, no. 3, pp. 361-366.

DESMOND, D., GALLAGHER, P., HENDERSON-SLATER, D. and CHATFIELD, R., 2008. Pain and psychosocial adjustment to lower limb amputation amongst prosthesis users. *Prosthetics and Orthotics International.* vol. 32, no. 2, pp. 244-252.

DEVLIN, M., PAULEY, T., HEAD, K. and GARFINKEL, S., 2004. Houghton Scale of prosthetic use in people with lower-extremity amputations: reliability, validity, and responsiveness to change. *Archives of Physical Medicine and Rehabilitation.* vol. 85, no. 8, pp. 1339-1344.

DEVLIN, N., SHAH, K. and FENG, Y., 2016. *Valuing Health-Related Quality of Life : An EQ-5D-5L Value Set for England.* London: Office of Health Ecconomics.

DILLINGHAM, T.R., PEZZIN, L.E. and SHORE, A.D., 2005. Reamputation, mortality, and health care costs among persons with dysvascular lower-limb amputations. *Archives of Physical Medicine & Rehabilitation.* 03, vol. 86, no. 3, pp. 480-486.

DOBSON, F., HINMAN, R.S., HALL, M., TERWEE, C.B., ROOS, E.M. and BENNELL, K.L., 2012. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. *Osteoarthritis and Cartilage.* vol. 20, no. 12, pp. 1548-1562.

DONOGHUE, D. and STOKES, E.K., 2009. How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *Journal of Rehabilitation Medicine.* vol. 41, no. 5, pp. 343-346.

EEKHOUT, I., DE VET, H.,C.W., TWISK, J.W.R., BRAND, J.P.L., DE BOER, M.,R. and HEYMANS, M.W., 2014. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology.* vol. 67, no. 3, pp. 335-342.

ENDERBY, P. and KEW, E., 1995. Outcome Measurement in Physiotherapy Using the World Health Organisation's Classification of Impairment, Disability and Handicap: A Pilot Study. *Physiotherapy.* vol. 81, no. 4, pp. 177-180.

FAVA, G.A., TOMBA, E. and SONINO, N., 2012. Clinimetrics: The science of clinical measurements. *International Journal of Clinical Practice.* vol. 66, no. 1, pp. 11-15.

FEYS, P., BIBBY, B., ROMBERG, A., SANTOYO, C., GEBARA, B., DE NOORDHOUT, B.M., KNUTS, K., BETHOUX, F., SKJERBAEK, A., JENSEN, E., BAERT, I., VANEY, C., DE GROOT, V. and DALGAS, U., 2014. Within-day variability on short and long walking tests in persons with multiple sclerosis. *Journal of the Neurological Sciences.* vol. 338, no. 1-2, pp. 183-187.

FISCHER, J.E., BACHMANN, L.M. and JAESCHKE, R., 2003. A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. *Intensive Care Medicine.* vol. 29, no. 7, pp. 1043-1051.

FLANSBJER, U.B., HOLMBACK, A.M., DOWNHAM, D., PATTEN, C. and LEXELL, J., 2005. Reliability of gait performance tests in men and women with hemiparesis after stroke. *Journal of Rehabilitation Medicine.* vol. 37, no. 2, pp. 75-82.

FORTINGTON, L., V., ROMMERS, G., M., GEERTZEN, J., H.B., POSTEMA, K. and DIJKSTRA, P., U., 2012. Mobility in Elderly People With a Lower Limb Amputation: A Systematic Review. *Journal of the American Medical Directors Association.* vol. 13, no. 4, pp. 319-325.

FRANCHIGNONI, F., BRUNELLI, S., ORLANDINI, D., FERRIERO, G. and TRABALLESI, M., 2003. Is the rivermead mobility index a suitable outcome measure in lower limb amputees? - A psychometric validation study. *Journal of Rehabilitation Medicine.* vol. 35, no. 3, pp. 141-144.

FRANCHIGNONI, F., GIORDANO, A., FERRIERO, G., ORLANDINI, D., AMORESANO, A. and PERUCCA, L., 2007a. Measuring mobility in people with lower limb amputation: Rasch analysis of the mobility section of the Prosthesis Evaluation Questionnaire. *Journal of Rehabilitation Medicine.* vol. 39, no. 2, pp. 138-144.

FRANCHIGNONI, F., MONTICONE, M., GIORDANO, A. and ROCCA, B., 2015. Rasch validation of the Prosthetic Mobility Questionnaire: A new outcome measure for assessing mobility in people with lower limb amputation. *Journal of Rehabilitation Medicine.* vol. 47, no. 5, pp. 460-465.

FRANCHIGNONI, F., GIORDANO, A., FERRIERO, G., MUÑOZ, S., ORLANDINI, D. and AMORESANO, A., 2007b. Rasch analysis of the Locomotor Capabilities Index-5 in people with lower limb amputation. *Prosthetics and Orthotics International.* vol. 31, no. 4, pp. 394-404.

FRANCHIGNONI, F., ORLANDINI, D., FERRIERO, G. and MOSCATO, T.A., 2004. Reliability, validity, and responsiveness of the locomotor capabilities index in adults with lower-limb amputation undergoing prosthetic training. *Archives of Physical Medicine and Rehabilitation.* vol. 85, no. 5, pp. 743-748.

FROSSARD, L., HAGBERG, K., HAGGSTROM, E., GOW, D.L., BRANEMARK, R. and PEARCY, M., 2010. Functional Outcome of Transfemoral Amputees Fitted With an Osseointegrated Fixation: Temporal Gait Characteristics. *Journal of Prosthetics and Orthotics.* vol. 22, no. 1, pp. 11-20.

FURLAN, A.D., MALMIVAARA, A., CHOU, R., MAHER, C.G., DEYO, R.A., SCHOENE, M., BRONFORT, G. and VAN TULDER, M.W., 2015. 2015 Updated Method Guideline for Systematic Reviews in the Cochrane Back and Neck Group. *Spine.* vol. 40, no. 21, pp. 1660-1673.

GAILEY, R.S., ROACH, K.E., APPLEGATE, E.B., CHO, B., CUNNIFFE, B., LICHT, S., MAGUIRE, M. and NASH, M.S., 2002. The Amputee Mobility Predictor: an instrument to assess determinants of the lower-limb amputee's ability to ambulate. *Archives of Physical Medicine and Rehabilitation.* vol. 83, no. 5, pp. 613-627.

GAILEY, R.S., GAUNAURD, I., AGRAWAL, V., BIOE, M.S., TOOLE, C.O., BIOE, M.S. and TOLCHIN, R., 2012. Application of self-report and performance-based outcome measures to determine functional differences between four categories of prosthetic feet. *Journal of Rehabilitation Research and Development.* vol. 49, no. 4, pp. 597-612.

GAILEY, R., ALLEN, K., CASTLES, J., KUCHARIK, J. and ROEDER, M., 2008. Review of secondary physical conditions associated with lower-limb amputation and long-term prosthesis use. *Journal of Rehabilitation Research and Development.* vol. 45, no. 1, pp. 15-29.

GALEA, M., 2005. Introducing Clinimetrics. *Australian Journal of Physiotherapy.* vol. 51, no. 3, pp. 139-140.

GALLAGHER, P., FRANCHIGNONI, F., GIORDANO, A. and MACLACHLAN, M., 2010. Trinity Amputation and Prosthesis Experience Scales: a psychometric assessment using classical test theory and Rasch analysis. *American Journal of Physical Medicine and Rehabilitation.* vol. 89, no. 6, pp. 487-496.

GALLAGHER, P., HORGAN, O., FRANCHIGNONI, F., GIORDANO, A. and MACLACHLAN, M., 2007. Body image in people with lower-limb amputation: a Rasch analysis of the Amputee Body Image Scale. *American Journal of Physical Medicine and Rehabilitation.* vol. 86, no. 3, pp. 205-215.

GALLAGHER, P. and MACLACHLAN, M., 2004. The Trinity Amputation and Prosthesis Experience Scales and quality of life in people with lower-limb amputation. *Archives of Physical Medicine & Rehabilitation.* vol. 85, no. 5, pp. 730-736.

GALLAGHER, P., O'DONOVAN, M., DOYLE, A. and DESMOND, D., 2011. Environmental barriers, activity limitations and participation restrictions experienced by people with major limb amputation. *Prosthetics and Orthotics International.* vol. 35, no. 3, pp. 278-284.

GARDINER, M.D., FAUX, S. and JONES, L.E., 2002. Inter-observer reliability of clinical outcome measures in a lower limb amputee population. *Disability and Rehabilitation.* vol. 24, no. 4, pp. 219-225.

GATCHEL, R.J. and MAYER, T.G., 2010. Testing minimal clinically important difference: consensus or conundrum? *The Spine Journal.* vol. 10, no. 4, pp. 321-327.

GAUNAURD, I., SPAULDING, S.E., AMTMANN, D., SALEM, R., GAILEY, R., MORGAN, S.J. and HAFNER, B.J., 2014. Use of and confidence in administering outcome measures among clinical prosthetists: Results from a national survey and mixed-methods training program. *Prosthetics and Orthotics International.* vol. 39, no. 4, pp. 314-321.

GAUTHIER-GAGNON, C. and GRISE, M., 1994. Prosthetic Profile of the Amputee questionnaire: validity and reliability. *Archives of Physical Medicine and Rehabilitation.* vol. 75, no. 12, pp. 1309-1314.

GAUTHIER-GAGNON, C., GRISE, M. and LEPAGE, Y., 1998. The Locomotor Capabilities Index: content validity. *Journal of Rehabilitation Outcomes Measurement.* vol. 2, no. 4, pp. 40-46.

GENIN, J.J., BASTIEN, G.J., FRANCK, B., DETREMBLEUR, C. and WILLEMS, P.A., 2008. Effect of speed on the energy cost of walking in unilateral traumatic lower limb amputees. *European Journal of Applied Physiology.* vol. 103, no. 6, pp. 655-663.

GIANOLA, S., GASPARINI, M., AGOSTINI, M., CASTELLINI, G., CORBETTA, D., GOZZER, P., LI, L.C., SIRTORI, V., TARICCO, M., TETZLAFF, J.M., TUROLLA, A., MOHER, D. and MOJA, L., 2013. Survey of the reporting characteristics of systematic reviews in rehabilitation. *Physical Therapy.* vol. 93, no. 11, pp. 1456-66.

GOKTEPE, A.S., CAKIR, B., YILMAZ, B. and YAZICIOGLU, K., 2010. Energy expenditure of walking with prostheses: Comparison of three amputation levels. *Prosthetics & Orthotics International.* vol. 34, no. 1, pp. 31-31.

GREEN, A., LILES, C., RUSHTON, A. and KYTE, D.G., 2014. Measurement properties of patient-reported outcome measures (PROMS) in Patellofemoral Pain Syndrome: A systematic review. *Manual Therapy.* vol. 19, pp. 517-526.

GREMEAUX, V., DAMAK, S., TROISGROS, O., FEKI, A., LAROCHE, D., PERENNOU, D., BENAIM, C. and CASILLAS, J., 2012. Selecting a test for the clinical assessment of balance and walking capacity at the definitive fitting state after unilateral amputation: a comparative study. *Prosthetics and Orthotics International.* vol. 36, no. 4, pp. 415-422.

GUYATT, G.H., DEYO, R.A., CHARLSON, M., LEVINE, M.N. and MITCHELL, A., 1989. Responsiveness and validity in health status measurement: a clarification. *Journal of Clinical Epidemiology.* vol. 42, no. 5, pp. 403-408.

GUYATT, G., WALTER, S. and NORMAN, G., 1987. Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases.* vol. 40, no. 2, pp. 171-178.

HAFNER, B.J. and SAWERS, A.B., 2016. Issues affecting the level of prosthetics research evidence : Secondary analysis of a systematic review. *Prosthetics and Orthotics International.* vol. 40, no. 1, pp. 31-43.

HAGBERG, K., TRANBERG, R., ZÜGNER, R. and DANIELSSON, A., 2011. Reproducibility of the Physiological Cost Index among Individuals with a Lower-Limb Amputation and Healthy Adults. *Physiotherapy Research International.* vol. 16, no. 2, pp. 92-100.

HALEY, S.M. and FRAGALA-PINKHAM, M., 2006. Interpreting change scores of tests and measures used in physical therapy. *Physical Therapy.* vol. 86, no. 5, pp. 735-743.

HAMAMURA, S., CHIN, T., KURODA, R., AKISUE, T., IGUCHI, T., KOHNO, H., KITAGAWA, A., TSUMURA, N. and KUROSAKA, M., 2009. Factors affecting prosthetic rehabilitation outcomes in amputees of age 60 years and over. *The Journal of International Medical Research.* vol. 37, no. 6, pp. 1921-1927.

HAMMOND, R., 2000. Evaluation of physiotherapy by measuring the outcome. *Physiotherapy.* vol. 86, no. 4, pp. 170-172.

HANCZAR, B., HUA, J., SIMA, C., WEINSTEIN, J., BITTNER, M. and DOUGHERTY, E.R., 2010. Small-sample precision of ROC-related estimates. *Bioinformatics.* vol. 26, no. 6, pp. 822-830.

HANSPAL, R.S., FISHER, K. and NIEVEEN, R., 2003. Prosthetic socket fit comfort score. *Disability & Rehabilitation.* vol. 25, no. 22, pp. 1278-1280.

HAWKINS, A.T., HENRY, A.J., CRANDELL, D.M. and NGUYEN, L.L., 2014. A systematic review of functional and quality of life assessment after major lower extremity amputation. *Annals of Vascular Surgery.* vol. 28, no. 3, pp. 763-780.

HAYS, R.D., FARIVAR, S.S. and LIU, H., 2005. Approaches and Recommendations for Estimating Minimally Important Differences for Health-Related Quality of Life Measures. *COPD: Journal of Chronic Obstructive Pulmonary Disease.* vol. 2, no. 1, pp. 63-67.

Health Improvement Scotland., 2012. *What is the clinical and cost effectiveness of microprocessor- controlled artificial knees compared with non-microprocessor-controlled alternatives? Evidence Note 44.* Glasgow: .

HEBERT, J.S., WOLFE, D.L., MILLER, W.C., DEATHE, A.B., DEVLIN, M. and PALLAVESHI, L., 2009. Outcome measures in amputation rehabilitation: ICF body functions. *Disability and Rehabilitation.* vol. 31, no. 19, pp. 1541-1554.

HEDEKER, D., GIBBONS, R. and WATERNAUX, C., 1999. Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics.* vol. 24, no. 1, pp. 70-93.

HEINEMANN, A.W., BODE, R.K. and O'REILLY, C., 2003. Development and measurement properties of the Orthotics and Prosthetics Users Survey (OPUS): A comprehensive set of clinical outcome instruments. *Prosthetics and Orthotics International.* vol. 27, no. 3, pp. 191-206.

HEINEMANN, A.W., CONNELLY, L., EHRLICH-JONES, L. and FATONE, S., 2014. Outcome Instruments for Prosthetics. *Physical Medicine and Rehabilitation Clinics of North America.* vol. 25, no. 1, pp. 179-198.

HERBERT, R., JAMTVEDT, G., MEAD, J. and HAGEN, K., 2005. Outcome measures measure outcomes, not effects of intervention. *The Australian Journal of Physiotherapy.* vol. 51, no. 1, pp. 3-4.

HIENGKAEW, V., JITAREE, K. and CHAIYAWAT, P., 2012. Minimal Detectable Changes of the Berg Balance Scale, Fugl-Meyer Assessment Scale, Timed “Up & Go” Test, Gait Speeds, and 2-Minute Walk Test in Individuals With Chronic Stroke With Different Degrees of Ankle Plantarflexor Tone. *Archives of Physical Medicine and Rehabilitation.* vol. 93, no. 7, pp. 1201-1208.

HIGHSMITH, M.J., KAHLE, J.T., MIRO, R.M. and MENGELKOCH, L.J., 2013. Ramp descent performance with the C-Leg and interrater reliability of the Hill Assessment Index. *Prosthetics and Orthotics International.* vol. 37, no. 5, pp. 362-368.

HILLMAN, S.J., DONALD, S.C., HERMAN, J., MCCURRACH, E., MCGARRY, A., RICHARDSON, A.M. and ROBB, J.E., 2010. Repeatability of a new observational gait score for unilateral lower limb amputees. *Gait and Posture.* vol. 32, no. 1, pp. 39-45.

HOENIG, H., LEE, J. and STINEMAN, M., 2010. Conceptual Overview of Frameworks for Measuring Quality in Rehabilitation. *Topics in Stroke Rehabilitation.* vol. 17, no. 4, pp. 239-251.

HOLMAN, N., YOUNG, R.J. and JEFFCOATE, W.J., 2012. Variation in the recorded incidence of amputation of the lower limb in England. *Diabetologia.* vol. 55, no. 7, pp. 1919-1925.

HOPEWELL, S., DUTTON, S., YU, L., CHAN, A. and ALTMAN, D.G., 2010. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *British Medical Journal (Clinical Research Ed.).* vol. 340, pp. c723-c723.

HORGAN, O. and MACLACHLAN, M., 2004. Psychosocial adjustment to lower-limb amputation: A review. *Disability and Rehabilitation.* vol. 26, no. 14-15, pp. 837-850.

HUANG, S., HSIEH, C., WU, R., TAI, C., LIN, C. and LU, W., 2011. Minimal detectable change of the timed "up & go" test and the dynamic gait index in people with parkinson disease. *Physical Therapy.* vol. 91, no. 1, pp. 114-121.

HUSTED, J., COOK, R., FAREWELL, V. and GLADMAN, D., 2000. Methods for assessing responsiveness:a critical review and recommendations. *Journal of Clinical Epidemiology.* vol. 53, pp. 459-468.

IYER, L.V., HALEY, S.M., WATKINS, M.P. and DUMAS, H.M., 2003. Establishing minimal clinically important differences for scores on the Pediatric Evaluation of Disability Inventory for inpatient rehabilitation. *Physical Therapy.* vol. 83, no. 10, pp. 888-898.

JAKOBSSON, U. and WESTERGREN, A., 2005. Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences.* vol. 19, no. 4, pp. 427-431.

JARL, G.M., HEINEMANN, A.W. and NORLING HERMANSSON, L.M., 2012. Validity evidence for a modified version of the Orthotics and Prosthetics Users Survey. *Disability and Rehabilitation: Assistive Technology.* vol. 7, no. 6, pp. 469-478.

JEFFCOATE, W., YOUNG, B. and HOLMAN, N., 2012. The variation in incidence of amputation throughout England. *Practical Diabetes.* vol. 29, no. 5, pp. 205-208.

JEROSCH-HEROLD, C., 2005. An evidence-based approach to choosing outcome measures: A checklist for the critical appraisal of validity, reliability and responsiveness studies. *British Journal of Occupational Therapy.* vol. 68, no. 8, pp. 347-353.

JETTE, D.U., HALBERT, J., IVERSON, C., MICELI, E. and SHAH, P., 2009. Use of standardized outcome measures in physical therapist practice: perceptions and applications. *Physical Therapy.* vol. 89, no. 2, pp. 125-135.

JOHNSTON, M.V. and GRAVES, D., 2008. SCoS-ASIA Boston meeting - Towards Guidelines for Evaluation of Measures: An Introduction with Application to Spinal Cord Injury. *Journal of Spinal Cord Medicine.* vol. 31, no. 1, pp. 13-26.

JOHNSTON, M.V. and DIJKERS, M.P., 2012. Toward improved evidence standards and methods for rehabilitation: Recommendations and challenges. *Archives of Physical Medicine and Rehabilitation.* vol. 93, no. 8, pp. S185-S199.

KAMPER, S., 2009. Global Rating of Change scales. *Australian Journal of Physiotherapy.* vol. 55, no. 4, pp. 289-289.

KANNENBERG, A., ZACHARIAS, B. and PROBSTING, E., 2014. Benefits of microprocessor-controlled prosthetic knees to limited community ambulators: Systematic review. *Journal of Rehabilitation Research and Development.* vol. 51, no. 10, pp. 1469-1496.

KARK, L., MCINTOSH, A.S. and SIMMONS, A., 2011. The use of the 6-min walk test as a proxy for the assessment of energy expenditure during gait in individuals with lower-limb amputation. *International Journal of Rehabilitation Research.* vol. 34, no. 3, pp. 227-234.

KIRBY, R.L., DUPUIS, D.J., MACPHEE, A.H., COOLEN, A.L., SMITH, C., BEST, K.L., NEWTON, A.M., MOUNTAIN, A.D., MACLEOD, D.A. and BONAPARTE, J.P., 2004. The Wheelchair Skills Test (version 2.4): measurement properties. *Archives of Physical Medicine & Rehabilitation.* vol. 85, no. 5, pp. 794-804.

KIRSHNER, B. and GUYATT, G., 1985. A methodological framework for assessing health indices. *Journal of Chronic Diseases.* vol. 38, no. 1, pp. 27-36.

KOHLER, F., CIEZA, A., STUCKI, G., GEERTZEN, J., BURGER, H., DILLON, M.P., SCHIAPPACASSE, C., ESQUENAZI, A., KISTENBERG, R.S. and KOSTANJSEK, N., 2009. Developing Core Sets for persons following amputation based on the International Classification of Functioning, Disability and Health as a way to specify functioning. *Prosthetics and Orthotics International.* vol. 33, no. 2, pp. 117-129.

KOHLER, F., XU, J., SILVA-WITHMORY, C. and AROCKIAM, J., 2011. Feasibility of using a checklist based on the International Classification of Functioning, Disability and Health as an outcome measure in individuals following lower limb amputation. *Prosthetics and Orthotics International.* vol. 35, no. 3, pp. 294-301.

KOTTNER, J., AUDIGE, L., BRORSON, S., DONNER, A., GAJEWSKI, B., J., HRO´BJARTSSON, A., ROBERTS, C., SHOUKRI, M. and STREINER, D., L., 2011. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies.* vol. 48, no. 6, pp. 661-671.

KRASNY-PACINI, A., HIEBEL, J., PAULY, F., GODON, S. and CHEVIGNARD, M., 2013. Goal Attainment Scaling in rehabilitation: A literature-based update. *Annals of Physical and Rehabilitation Medicine.* 04, vol. 56, no. 3, pp. 212-230.

KRISTENSEN, M.T., NIELSEN, A.Ø., TOPP, U.M., JAKOBSEN, B., NIELSEN, K.J., JUUL-LARSEN, H. and JAKOBSEN, T.L., 2014. Number of test trials needed for performance stability and interrater reliability of the one leg stand test in patients with a major non-traumatic lower limb amputation. *Gait and Posture.* vol. 39, no. 1, pp. 424-429.

LAFERRIER, J.Z. and GAILEY, R., 2010. Advances in lower-limb prosthetic technology. *Physical Medicine and Rehabilitation Clinics of North America.* vol. 21, no. 1, pp. 87-110.

LAHIRI, S. and GHOSH-DAS, P., 2012. Indian Journal of Physiotherapy and Occupational Therapy. *Indian Journal of Physiotherapy and Occupational Therapy.* vol. 6, no. 2, pp. 105-107.

LANDIS, J.R. and KOCH, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics.* vol. 33, no. 1, pp. 159-174.

LARSEN, C.M., JUUL-KRISTENSEN, B., LUND, H. and SØGAARD, K., 2014. Measurement properties of existing clinical assessment methods evaluating scapular positioning and function. A systematic review. *Physiotherapy Theory and Practice.* vol. 30, no. 7, pp. 453-482.

LEGRO, M.W., REIBER, G., DEL AGUILA, M., AJAX, M.J., BOONE, D.A., LARSEN, J.A., SMITH, D.G. and SANGEORZAN, B., 1999. Issues of importance reported by persons with lower limb amputations and prostheses. *Journal of Rehabilitation Research and Development.* vol. 36, no. 3, pp. 155-163.

LEMAIRE, E.D., 2013. *ISPO Recommendation for Defining Participants in Prosthetics Research.* ISPO Scientific Committee.

LEYLAND, A.H., 2005. Socioeconomic gradients in the prevalence of cardiovascular disease in Scotland: The roles of composition and context. *Journal of Epidemiology and Community Health.* vol. 59, no. 9, pp. 799-803.

LIANG, M.H., 1995. Evaluating measurement responsiveness. *The Journal of Rheumatology.* vol. 22, no. 6, pp. 1191-2.

LIN, S. and BOSE, N.H., 2008. Six-minute walk test in persons with transtibial amputation. *Archives of Physical Medicine and Rehabilitation.* vol. 89, no. 12, pp. 2354-2359.

MACEFIELD, R.C., JACOBS, M., KORFAGE, I.J., NICKLIN, J., WHISTANCE, R.N., BROOKES, S.T., SPRANGERS, M.A.G. and BLAZEBY, J.M., 2014. Developing core outcomes sets: methods for identifying and including patient-reported outcomes (PROs). *Trials.* vol. 15, no. 1, pp. 49-49.

MAJOR, M.J., FATONE, S. and ROTH, E.J., 2013. Validity and reliability of the berg balance scale for community-dwelling persons with lower-limb amputation. *Archives of Physical Medicine and Rehabilitation.* vol. 94, no. 11, pp. 2194-2202.

MCPHERSON, K., GIBSON, B. and LEPLEGE, A., 2015. *Rethinking Rehabilitation: Theory and practice.* Canada: Apple Academic Press.

MILLER, W.C., DEATHE, A.B. and HARRIS, J., 2004. Measurement properties of the Frenchay Activities Index among individuals with a lower limb amputation. *Clinical Rehabilitation.* vol. 18, no. 4, pp. 414-422.

MILLER, W.C., DEATHE, A.B. and SPEECHLEY, M., 2003. Psychometric properties of the Activities-Specific Balance Confidence Scale among individuals with a lower-limb amputation. *Archives of Physical Medicine and Rehabilitation.* vol. 84, no. 5, pp. 656-661.

MILLER, W.C., DEATHE, A.B. and SPEECHLEY, M., 2001. Lower extremity prosthetic mobility: a comparison of 3 self-report scales. *Archives of Physical Medicine and Rehabilitation.* vol. 82, no. 10, pp. 1432-1440.

MOHER, D., LIBERATI, A., TETZLAFF, J., ALTMAN, D.G. and GRP, P., 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement (Reprinted from Annals of Internal Medicine). *Physical Therapy.* vol. 89, no. 9, pp. 873-880.

MOKKINK, L.B., TERWEE, C.B., KNOL, D.L., BOUTER, L.M., DE VET, H.C.W., STRATFORD, P.W., ALONSO, J. and PATRICK, D.L., 2006. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BioMed Central Medical Research Methodology.* vol. 6, no. 1, pp. 2-9.

MOKKINK, L.B., TERWEE, C.B., PATRICK, D.L., ALONSO, J., STRATFORD, P.W., KNOL, D.L., BOUTER, L.M. and DE VET, H., 2010a. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research.* vol. 19, no. 4, pp. 539-549.

MOKKINK, L.B., TERWEE, C.B., PATRICK, D.L., ALONSO, J., STRATFORD, P.W., KNOL, D.L., BOUTER, L.M. and DE VET, H., 2010b. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology.* vol. 63, no. 7, pp. 737-745.

MOKKINK, L.B., TERWEE, C.B., KNOL, D.L., STRATFORD, P.W., ALONSO, J., PATRICK, D.L., BOUTER, L.M. and DE VET, H.,C.W., 2010c. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BioMedCentral Medical Research Methodology.* vol. 10, no. 1, pp. 22-22.

MOKKINK, L.B., TERWEE, C.B., PATRICK, D.L., ALONSO, J., STRATFORD, P.W., KNOL, D.L., BOUTER, L.M.and DE VET, H.C., 2012. *The COSMIN checklist manual.*

MONTEIRO, R.P.D.A., PFEIFER, L.I., SOARES, I. and Santos,Alex de Assis dos Santos, 2013. Validation of the functional and social performance - DSF-84 checklist: preliminary study. *Disability and Rehabilitation.* vol. 35, no. 18, pp. 1527-1533.

MORRIS, R.W., WHINCUP, P.H., LAMPE, F.C., WALKER, M., WANNAMETHEE, S.G. and SHAPER, A.G., 2001. Geographic variation in incidence of coronary heart disease in Britain: the contribution of established risk factors. *Heart (British Cardiac Society).* vol. 86, no. 3, pp. 277-83.

MORSE, B.C., CULL, D.L., KALBAUGH, C., CASS, A.L. and TAYLOR, S.M., 2008. Through-knee amputation in patients with peripheral arterial disease: a review of 50 cases. *Journal of Vascular Surgery.* vol. 48, no. 3, pp. 638-643.

MOSELEY, A.M., HERBERT, R.D., SHERRINGTON, C. and MAHER, C.G., 2002. Evidence for physiotherapy practice: a survey of the Physiotherapy Evidence Database (PEDro). *The Australian Journal of Physiotherapy.* vol. 48, no. 1, pp. 43-49.

MOXEY, P.W., HOFMAN, D., HINCHLIFFE, R.J., JONES, K., THOMPSON, M.M. and HOLT, P.J.E., 2010. Epidemiological study of lower limb amputation in England between 2003 and 2008. *The British Journal of Surgery.* vol. 97, no. 9, pp. 1348-1353.

MOXEY, P.W., GOGALNICEANU, P., HINCHLIFFE, R.J., LOFTUS, I.M., JONES, K.J., THOMPSON, M.M. and HOLT, P.J., 2011. Lower extremity amputations - a review of global variability in incidence. *Diabetic Medicine.* vol. 28, no. 10, pp. 1144-1153.

MUNIN, M.C., ESPEJO-DE GUZMAN, M., BONINGER, M.L., FITZGERALD, S.G., PENROD, L.E. and SINGH, J., 2001. Predictive factors for successful early prosthetic ambulation among lower-limb amputees. *Journal of Rehabilitation Research and Development.* vol. 38, no. 4, pp. 379-384.

MURRAY, C.D. and FORSHAW, M.J., 2013. The experience of amputation and prosthesis use for adults: a metasynthesis. *Disability and Rehabilitation.* vol. 35, no. 14, pp. 1133-42.

NORMAN, G.R., SLOAN, J.A. and WYRWICH, K.W., 2003. Interpretation of changes in health-related quality of life the remarkable universality of half a standard deviation. *Medical Care.* vol. 41, no. 5, pp. 582-592.

NORMAN, G.R., STRATFORD, P. and REGEHR, G., 1997. Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *Journal of Clinical Epidemiology.* vol. 50, no. 8, pp. 869-879.

O'CONNELL, N.,E., MOSELY, G.L., MCAULEY, J., WAND, B. and HERBERT, R.D., 2015. Interpreting Effectiveness Evidence in Pain: Short Tour of Contemporary Issues. *Physical Therapy.* vol. 95, no. 8, pp. 1087-1094.

ORENDURFF, M.S., 2012. Literature Review of Published Research Investigating Microprocessor-Controlled Prosthetic Knees: 2010-2012. *Journal of Prosthetic and Orthotics.* vol. 35, no. 25, pp. 41-46.

PANESAR, B.S., MORRISON, P. and HUNTER, J., 2001. A comparison of three measures of progress in early lower limb amputee rehabilitation. *Clinical Rehabilitation.* vol. 15, no. 2, pp. 157-171.

PARKER, K., KIRBY, R.L., ADDERSON, J. and THOMPSON, K., 2010. Ambulation of people with lower-limb amputations: relationship between capacity and performance measures. *Archives of Physical Medicine & Rehabilitation.* vol. 91, no. 4, pp. 543-549.

PIN, T.W., 2014. Psychometric Properties of 2-Minute Walk Test: A Systematic Review. *Archives of Physical Medicine & Rehabilitation.* vol. 95, no. 9, pp. 1759-1775.

PODSIADLO, D. and RICHARDSON, S., 1991. The Timed Up & Go• : A Test of Basic Functional Mobility for Frail Elderly Persons. *Journal of the American Geriatrics Society.* vol. 39, no. 2, pp. 142-148.

POLKEY, M.I., SPRUIT, M.A., EDWARDS, L.D., WATKINS, M.L., PINTO-PLATA, V., VESTBO, J., CALVERLEY, P.M.A., TAL-SINGER, R., AGUSTI, A., BAKKE, P.S., COXSON, H.O., LOMAS, D.A., MACNEE, W., RENNARD, S., SILVERMAN, E.K., MILLER, B.E., CRIM, C., YATES, J., WOUTERS, E.F.M. and CELLI, B., 2013. Six-minute-walk test in chronic obstructive pulmonary disease: Minimal clinically important difference for death or hospitalization. *American Journal of Respiratory and Critical Care Medicine.* vol. 187, no. 4, pp. 382-386.

PRINSEN, C.A.C., VOHRA, S., ROSE, M.R., KING-JONES, S., ISHAQUE, S., BHALOO, Z., ADAMS, D. and TERWEE, C.B., 2014. Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. *Trials.* 07, vol. 15, no. 1, pp. 247.

RANKIN, G. and STOKES, M., 1998. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical Rehabilitation.* vol. 12, no. 3, pp. 187-199.

RAYA, M., A., GAILEY, R., S., GAUNAURD, I., A., GANYARD, H., KNAPP-WOOD, J., MCDONOUGH, K. and PALMISANO, T., 2013. Amputee Mobility Predictor-Bilateral: A performance-based measure of mobility for people with bilateral lower-limb loss. *Journal of Rehabilitation Research and Development.* vol. 50, no. 7, pp. 961-968.

REID, L., THOMSON, P., BESEMANN, M. and DUDEK, N., 2015. Going places: Does the two-minute walk test predict the six-minute walk test in lower extremity amputees? *Journal of Rehabilitation Medicine.* vol. 47, no. 3, pp. 256-261.

RESNIK, L. and BORGIA, M., 2011. Reliability of outcome measures for people with lower-limb amputations: distinguishing true change from statistical error. *Physical Therapy.* vol. 91, no. 4, pp. 555-565.

REVICKI, D., HAYS, R.D., CELLA, D. and SLOAN, J., 2008. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology.* vol. 61, no. 2, pp. 102-109.

ROACH, K., 2006. Measurement of Health Outcomes: Reliability, Validity and Responsiveness. *Journal of Prosthetic and Orthotics.* vol. 18, no. 1, pp. 8-12.

ROMMERS, G.M., RYALL, N.H., KAP, A., DE LAAT, F. and VAN, D.L., 2008. The mobility scale for lower limb amputees: The SIGAM/WAP mobility scale. *Disability and Rehabilitation: An International, Multidisciplinary Journal.* vol. 30, no. 15, pp. 1106-1115.

ROMMERS, G.M., VOS, L., GROOTHOFF, J.W. and EISMA, W.H., 2001. Mobility of people with lower limb amputations: scales and questionnaires: a review. *Clinical Rehabilitation.* vol. 15, no. 1, pp. 92-102.

ROSS, M., 1989. Relation of implicit theories to the construction of personal histories. *Psychological Review.* vol. 96, no. 2, pp. 341-357.

RUSHTON, P.W. and MILLER, W.C., 2002. Goal Attainment Scaling in the rehabilitation of patients with lower-extremity amputations: a pilot study. *Archives of Physical Medicine and Rehabilitation.* vol. 83, no. 6, pp. 771-775.

RUSHTON, P.W., MILLER, W.C. and DEATHE, A.B., 2015. Minimal clinically important difference of the L Test for individuals with lower limb amputation: A pilot study. *Prosthetics and Orthotics International.* vol. 39, no. 6, pp. 470-476.

RYALL, N.H., EYRES, S.B., NEUMANN, V.C., BHAKTA, B.B. and TENNANT, A., 2003a. Is the Rivermead Mobility Index appropriate to measure mobility in lower limb amputees? *Disability and Rehabilitation.* vol. 25, no. 3, pp. 143-153.

RYALL, N.H., EYRES, S.B., NEUMANN, V.C., BHAKTA, B.B. and TENNANT, A., 2003b. The SIGAM mobility grades: a new population-specific measure for lower limb amputees. *Disability and Rehabilitation.* vol. 25, no. 15, pp. 833-844.

RYAN, R.M. and DECI, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist.* vol. 55, no. 1, pp. 68-78.

SAKAKIBARA, B.M., MILLER, W.C. and BACKMAN, C.L., 2011. Rasch analyses of the Activities-specific Balance Confidence Scale with individuals 50 years and older with lower-limb amputations. *Archives of Physical Medicine and Rehabilitation.* vol. 92, no. 8, pp. 1257-1263.

SALTER, K., JUTAI, J.W., TEASELL, R., FOLEY, N.C., BITENSKY, J. and BAYLEY, M., 2005. Issues for selection of outcome measures in stroke rehabilitation: ICF Participation. *Disability & Rehabilitation.* vol. 27, no. 9, pp. 507-528.

SANSAM, K., NEUMANN, V., O'CONNOR, R. and BHAKTA, B., 2009. Predicting walking ability following lower limb amputation: a systematic review of the literature. *Journal of Rehabilitation Medicine.* vol. 41, no. 8, pp. 593-603.

SAWERS, A.B. and HAFNER, B.J., 2013. Outcomes associated with the use of microprocessor-controlled prosthetic knees among individuals with unilateral transfemoral limb loss: A systematic review. *The Journal of Rehabilitation Research and Development.* vol. 50, no. 3, pp. 273-273.

SCHAFER, J.L. and GRAHAM, J.W., 2002. Missing data: Our view of the state of the art. *Psychological Methods.* vol. 7, no. 2, pp. 147-177.

SCHAFFALITZKY, E., GALLAGHER, P., MACLACHLAN, M. and RYALL, N., 2011. Understanding the benefits of prosthetic prescription: exploring the experiences of practitioners and lower limb prosthetic users. *Disability and Rehabilitation.* vol. 33, no. 15-16, pp. 1314-1323.

SCHAFFALITZKY, E., GALLAGHER, P., MACLACHLAN, M. and WEGENER, S.T., 2012. Developing consensus on important factors associated with lower limb prosthetic prescription and use. *Disability and Rehabilitation.* vol. 34, no. 24, pp. 2085-2094.

SCHOLTES, V.A., TERWEE, C.B. and POOLMAN, R.W., 2011. What makes a measurement instrument valid and reliable? *Injury.* vol. 42, no. 3, pp. 236-240.

SCHOPPEN, T., BOONSTRA, A., GROOTHOFF, J.W., DE VRIES, J., GOEKEN, L. and EISMA, W.H., 1999. The Timed "up and go" test: reliability and validity in persons with unilateral lower limb amputation. *Archives of Physical Medicine and Rehabilitation.* vol. 80, no. 7, pp. 825-828.

SCHOPPEN, T., BOONSTRA, A., GROOTHOFF, J.W., DE VRIES, J., GOEKEN, L.N. and EISMA, W.H., 2003. Physical, mental, and social predictors of functional outcome in unilateral lower-limb amputees. *Archives of Physical Medicine and Rehabilitation.* vol. 84, no. 6, pp. 803-811.

SCHULZ, K.F., ALTMAN, D.G. and MOHER, D., 2010. CONSORT 2010 Statement : Updated Guidelines for Reporting Parallel Group Randomized Trials. *Annals of Internal Medicine.* vol. 1996, no. 14, pp. 727-732.

SCOTT, H., PATEL, R. and HEBENTON, J., 2016. *A Survey of the Lower Limb Amputee Population in Scotland 2013.* Glasgow: Scottish Physiotherapy Amputee Research Group (SPARG).

SCRIVENER, K., SHERRINGTON, C. and SCHURR, K., 2013. A systematic review of the responsiveness of lower limb physical performance measures in inpatient care after stroke. *BioMedCentral Neurology.* vol. 13, no. 1, pp. 4.

SHAMSEER, L., MOHER, D., CLARKE, M., GHERSI, D., LIBERATI, A., PETTICREW, M., SHEKELLE, P. and STEWART, L.a., 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *British Medical Journal.* vol. 349, no. 1, pp. g7647.

SHROUT, P.E. and FLEISS, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin.* vol. 86, no. 2, pp. 420-428.

SHULTZ, S. and OLSZEWSKI, A., 2013. A sytematic review of outcome tools used to measure lower leg conditions. *International Journal of Sports Physical Therapy.* vol. 8, no. 6, pp. 838-848.

SINHA, R. and WIM J., A., 2011. A systematic literature review of quality of life in lower limb amputees. *Disability and Rehabilitation.* vol. 33, no. 11, pp. 883-899.

SKINNER, A. and TURNER-STOKES, L., 2006. The use of standardized outcome measures in rehabilitation centres in the UK. *Clinical Rehabilitation.* vol. 20, no. 7, pp. 609-615.

STEENBEEK, D., KETELAAR, M., GALAMA, K. and GORTER, J.W., 2008. Goal attainment scaling in paediatric rehabilitation: A report on the clinical training of an interdisciplinary team. *Child: Care, Health and Development.* vol. 34, no. 4, pp. 521-529.

STEFFEN, T. and SENEY, M., 2008. Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-Item Short-Form Health Survey, and the Unified Parkinson Disease Rating Scale in people with parkinsonism [corrected] [published erratum appears in PHYS THER 2010 Mar;90(3):462]. *Physical Therapy.* vol. 88, no. 6, pp. 733-746.

STEINER, W.A., RYSER, L., HUBER, E., UEBELHART, D., AESCHLIMANN, A. and STUCKI, G., 2002. Use of the ICF model as a clinical problem-solving tool in physical therapy and rehabilitation medicine. *Physical Therapy.* vol. 82, no. 11, pp. 1098-1107.

STOKES, E.K., 2010. *Rehabilitation Outcome Measures.* Edinburgh: Churchill Livingstone.

STOLWIJK-SWUSTE, J., BEELEN, A., LANKHORST, G.J., NOLLET, F., DEKKER, J., VAN DIJK, G.M., VAN DEN ENDE, C.,H.M., POST, B., DE HAAN, R.J. and SPEELMAN, H., 2008. SF36 physical functioning scale and 2-minute walk test advocated as core qualifiers to evaluate physical functioning in patients with late-onset sequelae of poliomyelitis. *Journal of Rehabilitation Medicine.* vol. 40, no. 5, pp. 387-392.

STRATFORD, P.W., 2004. Getting more from the literature: estimating the standard error of measurement from reliability studies. *Physiotherapy Canada.* vol. 56, no. 1, pp. 27-30.

STRATFORD, P.W., BINKLEY, J., SOLOMON, P., FINCH, E., GILL, C. and MORELAND, J., 1996a. Defining the minimum level of detectable change for the Roland-Morris Questionnaire... including commentary by Riddle DL with author response. *Physical Therapy.* vol. 76, no. 4, pp. 359-368.

STRATFORD, P.W., BINKLEY, J.M. and RIDDLE, D.L., 1996b. Health status measures: strategies and analytic methods for assessing change scores. *Physical Therapy.* vol. 76, no. 10, pp. 1109-1123.

STRATFORD, P.W., BINKLEY, J.M., RIDDLE, D.L. and GUYATT, G.H., 1998. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Physical Therapy.* vol. 78, no. 11, pp. 1186-96.

STREINER, D.L., 2003. Clinimetrics vs. psychometrics: an unnecessary distinction. *Journal of Clinical Epidemiology.* vol. 56, no. 12, pp. 1142-1145.

STREINER, D.L., NORMAN, G.R. and CAIRNEY, J., 2014. *Health measurement scales: A practical guide to their development and use.* United Kingdom: Oxford University Press.

STUCKI, G., DALTROY, J., JOHANNESSON, M. and LIANG, M., 1996. Interpretation of Change Scores in Ordinal Clinical Scales and Health Status Measures: The Whole May Not Equal the Sum of the Parts. *Journal of Clinical Epidemiology.* vol. 49, no. 7, pp. 711-717.

SULLIVAN, J., UDEN, M., ROBINSON, K.P. and SOORIAKUMARAN, S., 2003. Rehabilitation of the trans-femoral amputee with an osseointegrated prosthesis: the United Kingdom experience. *Prosthetics and Orthotics International.* vol. 27, no. 2, pp. 114-120.

SVENSSON, E., 2001. Guidelines to statistical evaluation of data from rating scales and questionnaires. *Journal of Rehabilitation Medicine.* vol. 33, no. 1, pp. 47-48.

TAVAKOL, M. and DENNICK, R., 2011. Making sense of Cronbach's alpha. *International Journal of Medical Education.* vol. 2, pp. 53-55.

TAYLOR, S.M., KALBAUGH, C.A., BLACKHURST, D.W., HAMONTREE, S.E., CULL, D.L., MESSICH, H.S., ROBERTSON, R.T., LANGAN, E.M., YORK, J.W., CARSTEN, C.G., SNYDER, B.A., JACKSON, M.R. and YOUKEY, J.R., 2005. Preoperative clinical factors predict postoperative functional outcomes after major lower limb amputation: an analysis of 553 consecutive patients. *Journal of Vascular Surgery.* vol. 42, no. 2, pp. 227-235.

TENNANT, A., 2007. Goal attainment scaling: current methodological challenges. *Disability and Rehabilitation.* vol. 29, no. 20-21, pp. 1583-1588.

TERWEE, C.B., DEKKER, F.W., WIERSINGA, W.M., PRUMMEL, M.F. and BOSSUYT, P.M.M., 2003. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research.* vol. 12, no. 4, pp. 349-362.

TERWEE, C.B., JANSMA, E.P., RIPHAGEN, I. and DE VET, H., 2009. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research.* vol. 18, no. 8, pp. 1115-1123.

TERWEE, C.B., MOKKINK, L.B., KNOL, D.L., OSTELO, R.W., BOUTER, L.M. and DE VET, H., 2012. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research.* vol. 21, no. 4, pp. 651-657.

TERWEE, C.B., BOT, S.D.M., DE BOER, M.,R., VAN DER WINDT, D.,A.W.M., KNOL, D.L., DEKKER, J., BOUTER, L.M. and DE VET, H.,C.W., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology.* vol. 60, no. 1, pp. 34-42.

TESIO, L., 2003. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine.* vol. 35, no. 3, pp. 105-115.

THEEVEN, P., HEMMEN, B., RINGS, F., MEYS, G., BRINK, P., SMEETS, R. and SEELEN, H., 2011. Functional added value of microprocessor-controlled prosthetic knee joints in daily life performance of Medicare Functional Classification Level-2 amputees. *Journal of Rehabilitation Medicine.* vol. 43, no. 10, pp. 906-915.

THEEVEN, P., HEMMEN, B., STEVENS, C., ILMER, E., BRINK, P. and SEELEN, H., 2010. Feasibility of a new concept for measuring actual functional performance in daily life of transfemoral amputees. *Journal of Rehabilitation Medicine.* vol. 42, no. 8, pp. 744-751.

TILSON, J.K., SULLIVAN, K.J., CEN, S.Y., ROSE, D.K., KORADIA, C.H., AZEN, S.P. and DUNCAN, P.W., 2010. Meaningful Gait Speed Improvement During the First 60 Days Poststroke: Minimal Clinically Important Difference. *Physical Therapy.* vol. 90, no. 2, pp. 196-208.

TUGWELL, P., BOERS, M., BROOKS, P., SIMON, L., STRAND, V. and IDZERDA, L., 2007. OMERACT: An international initiative to improve outcome measurement in rheumatology. *Trials.* vol. 8, no. 1, pp. 38-38.

TURNER, D., SCHUNEMANN, H.J., GRIFFITH, L.E., BEATON, D.E., GRIFFITHS, A.M., CRITCH, J.N. and GUYATT, G.H., 2009. Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. *Journal of Clinical Epidemiology.* vol. 62, no. 4, pp. 374-379.

TURNER-STOKES, L. and TURNER-STOKES, T., 1997. The use of standardized outcome measures in rehabilitation centres in the UK. *Clinical Rehabilitation.* vol. 11, no. 4, pp. 306-313.

TURNER-STOKES, L., 2009. Goal attainment scaling (GAS) in rehabilitation: a practical guide. *Clinical Rehabilitation.* vol. 23, no. 4, pp. 362-370.

TYSON, S., 2009. The psychometric properties and clinical utility of measures of walking and mobility in neurological conditions: a systematic review. *Clinical Rehabilitation.* vol. 23, pp. 1018-1033.

UNIPOD., 2013. *Limbless Statistics Annual Report 2011-12.* United National Institute for Prosthetics, & Orthotics Development ed., Manchester: University of Salford.

USTUN, T.B., CHATTERJI, S., BICKENBACH, J., KOSTANJSEK, N. and SCHNEIDER, M., 2003. The International Classification of Functioning, Disability and Health: a new tool for understanding disability and health. *Disability and Rehabilitation.* vol. 25, no. 11-12, pp. 565-571.

VAN KAMPEN, D.A., WILLEMS, W.J., van Beers, Loes W. A. H., CASTELEIN, R.M., SCHOLTES, V.A.B. and TERWEE, C.B., 2013. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *Journal of Orthopaedic Surgery & Research.* vol. 8, no. 1, pp. 40.

VAN TULDER, M., FURLAN, a., BOMBARDIER, C. and BOUTER, L., 2003. Updated method guidelines for systematic reviews in the cochrane collaboration. *Spine.* vol. 28, no. 12, pp. 1290-1299.

VAN TWILLERT, S., STUIVE, I., GEERTZEN, J., POSTEMA, K. and LETTINGA, A., 2014. Functional performance, participation and autonomy after discharge from prosthetic rehabilitation: Barriers, facilitators and outcomes. *Journal of Rehabilitation Medicine.* vol. 46, no. 9, pp. 915-923.

VAN VELZEN, J., VAN BENNEKOM, C., POLOMSKI, W., SLOOTMAN, J.R., LH and HOUDIJK, H., 2006. Physical capacity and walking ability after lower limb amputation: a systematic review. *Clinical Rehabilitation.* vol. 20, no. 11, pp. 999-1016.

VON ELM, E., ALTMAN, D.G., EGGER, M., POCOCK, S.J., GOTZSCHE, P.C. and VANDENBROUCKE, J.P., 2014. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *International Journal of Surgery.* vol. 12, no. 12, pp. 1495-1499.

WALTERS, S.J. and BRAZIER, J.E., 2005. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life Research.* vol. 14, no. 6, pp. 1523-1532.

WARE, J.E. and GANDEK, B., 1998. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *Journal of Clinical Epidemiology.* vol. 51, no. 11, pp. 903-912.

WEIR, J.P., 2005. Quantifying Test Re-test Reliability Using The Intraclass Correlation Coefficient And The SEM. *Journal of Strength and Conditioning Research.* vol. 19, no. 1, pp. 231-40.

WHO., (2002), World Health Organisation International Classification of Functioning, Disability and Health (ICF)., WHO: Geneva

WILLIAMSON, P., ALTMAN, D., BLAZEBY, J., CLARKE, M. and GARGON, E., 2012. Driving up the quality and relevance of research through the use of agreed core outcomes. *Journal of Health Services Research & Policy.* vol. 17, no. 1, pp. 1-2.

WILLIAMSON, P.R., ALTMAN, D.G., BLAZEBY, J.M., CLARKE, M. and GARGON, E., 2011. The COMET (Core Outcome Measures in Effectiveness Trials) Initiative. *Trials.* vol. 12, pp. A70-A70.

WILLIAMSON, P.R., ALTMAN, D.G., BLAZEBY, J.M., CLARKE, M., DEVANE, D., GARGON, E. and TUGWELL, P., 2012. Developing core outcome sets for clinical trials: issues to consider. *Trials.* vol. 13, pp. 1-1.

WITTINK, H., ROGERS, W., SUKIENNIK, A. and CARR, D.B., 2003. Physical functioning: self-report and performance measures are related but distinct. *Spine.* vol. 28, no. 20, pp. 2407-2413.

WONG, C.K., 2014. Interrater Reliability of the Berg Balance Scale for People With Lower Limb Amputations. *Physical Therapy.* vol. 94, no. 3, pp. 371-378.

World Health Organisation., September 2003. *ICF Checklist.* Available from: http://www.who.int/classifications/icf/training/icfchecklist.pdf.

XU, J., KOHLER, F. and DICKSON, H., 2011. Systematic review of concepts measured in individuals with lower limb amputation using the International Classification of Functioning, Disability and Health as a reference. *Prosthetics and Orthotics International.* vol. 35, no. 3, pp. 262-268.

YOUNG, J., ROWLEY, L., LALOR, S., CODY, C. and WOOLLEY, H., 2015. *Measuring Change: An Introduction to Clinical Outcome Measures in Prosthetics and Orthotics.* Paisley: British Association of Prosthetists and Orthotists.

Appendices

Appendix 1

Allied health professionals survey questionnaire

Queen Margaret University
EDINBURGH

What are you measuring and what matters most?

Please complete the following questions by ticking the most applicable box. If you do not wish to complete a question then leave it blank and move onto the next one.

Please remember to fill the survey in on your own to keep your answers anonymous.

Question 1: What is your Profession, please tick:

Occupational Therapist	
Physiotherapist	
Prosthetist	

Question 2: How many years have you been qualified, please tick:

Up to 5yrs	
5-10 yrs	
>10yrs	

Question 3: How many years have you working with amputees, please tick:

Up to 5yrs	
5-10 yrs	
>10yrs	

Question 4: What % of your work is connected with amputee rehabilitation, please tick:

0 - 30%	
31 - 60%	
61 – 100%	

Question 5: Which members of the Multi-disciplinary Team do you regularly work with, i.e. on a daily basis, please tick all that apply:

Clinical Psychologist	
Doctor	
Nurse	
Prosthetist	
Physiotherapist	
Occupational Therapist	

Queen Margaret University
EDINBURGH

Question 6: Please tick the Outcome Measures you use regularly, i.e. at least once a week with the following amputee populations:

	Activity Level K0 – K2 Non-limb wearer, indoor or limited outdoor ambulator		**Activity Level K3** Wearing limb daily and fully ambulant outdoors		**Activity Level K4** Undertakes athletic activities in addition to daily tasks	
	Primary	**Established**	**Primary**	**Established**	**Primary**	**Established**
Activities-specific Balance Confidence Scale-UK						
Amputation Related Body Image Scale						
Amputee Activity Score						
Amputee Mobility Predictor with Prosthesis						
Attitude to Artificial Limb Questionnaire						
Barthel Index						
Body Image Questionnaire						
COPM						
Frenchay Activities Index (FAI)						
Functional Measure for Amputees (FMA)						
Functional Independence Measure (FIM)						
Houghton Scale						
Locomotor Capabilities Index						
LCI-5						
Orthotics and Prosthetics National Outcomes Tool						
Patient Generated Index						
Perceived Social Stigma Scale						
Prosthesis Evaluation Questionnaire						
Prosthetic Profile of the Amputee						
Prosthetics National Outcomes Tool (OPOT)						
Prosthetic Observational Gait Score (POGS)						
Rivermead Mobility Index (RMI)						
Russek's Code						

Queen Margaret University
EDINBURGH

	Activity Level K0 – K2 Non-limb wearer, indoor or limited outdoor ambulator		Activity Level K3 Wearing limb daily and fully ambulant outdoors		Activity Level K4 Undertakes athletic activities in addition to daily tasks	
	Primary	**Established**	**Primary**	**Established**	**Primary**	**Established**
Short Form 12 or 36 Health Survey (SF-12 or 36)						
Sickness Impact Profile (SIP)						
SIGAM						
Socket Comfort Score						
Timed "Up and Go" Test (TUG)						
Timed Walk Test (2min or 6min)						
Trinity Amputation and Prosthesis Experience Scales						

Please add any other outcome measure not listed that you use regularly (at least once a week)

_____

_____

_____

_____

Question 7: Did you use, then stop using any of the Outcome Measures listed above? please delete as appropriate. Yes /No

If Yes, which one(s) _____
and what was the reason, please tick all that apply:

Lack of time	
No relevance to clinical practice	
Lack of sensitivity, i.e. ceiling or floor effect observed	
Other, please state:	

Queen Margaret University
EDINBURGH

Question 8: What areas, from the following list, would you consider are the most important to achieving a successful rehabilitation with a prosthesis when working with the following amputee populations.

Please rank in order with 1=most important and 10=least important in each group.

	Activity Level K0 – K2 Non-limb wearer, indoor or limited outdoor ambulator		**Activity Level K3** Wearing limb daily and fully ambulant outdoors		**Activity Level K4** Undertakes some athletic activities in addition to daily tasks	
	Primary	**Established**	**Primary**	**Established**	**Primary**	**Established**
Balance & Coordination						
Energy, Drive & Motivation						
General Muscle strength						
General Range of Movement						
Manual Dexterity						
Memory						
Pain Control						
Self Confidence						
Socket fit & comfort						
Support of family & friends						

Please add any other area not listed that you consider important, and for which group

_____

_____

_____

_____

THANK YOU

Please send the completed survey back to me using the stamped addressed envelope.

Appendix 2

Lower limb amputee survey questionnaire

Queen Margaret University
EDINBURGH

What matters most?

Please complete the following questions by ticking the most applicable box. If you do not wish to complete a question then leave it blank and move onto the next one.

Please remember to fill the survey in on your own to keep your answers anonymous.

Question 1: What is your age range, please tick:

18-25	
26-50	
51-65	
>65	

Question 2: Are you, please tick:

Male	
Female	

Question 3: What is the level of your amputation(s), please tick all that are applicable:

	One leg	Both legs
Below-knee		
Through-knee		
Above-knee		
Other, please state:		

Question 4: How many years have you had a prosthesis, please tick:

1 to 5yrs	
5-10 yrs	
>10yrs	

Queen Margaret University
EDINBURGH

Question 5: What factors, from the following list, would you consider were the most important to your rehabilitation with your prosthetic limb(s)?

Please rank in order with 1 = the most important and 10 = the least important.

Balance & Coordination	
Energy, Drive & Motivation	
General Muscle strength	
General Range of Movement	
Manual Dexterity	
Memory	
Pain Control	
Self Confidence	
Socket fit & comfort	
Support of family & friends	

Please add any other factor that is not listed that you considered was important to your rehabilitation

_____

_____

_____

_____

THANK YOU

Please send the completed survey back to me using the stamped addressed envelope.

Appendix 3

Number of Allied Health Professional respondents who regularly used the listed outcome measures, by profession

Physiotherapists' responses (n=23)

	Activity Level K0 – K2 Non-limb wearer, indoor or limited outdoor ambulator		**Activity Level K3** Wearing limb daily and fully ambulant outdoors		**Activity Level K4** Undertakes athletic activities in addition to daily tasks		**TOTALS**
	Primary	**Established**	**Primary**	**Established**	**Primary**	**Established**	
Activities-specific Balance Confidence (ABC) Scale-UK	3	1	3	2	1	2	**12**
Amputee Activity Score	1	1	1	1	1	1	**6**
Amputee Mobility Predictor with Prosthesis	4	2	4	2	2	1	**15**
Attitude to Artificial Limb Questionnaire							
Barthel Index	2		2		2		**6**
Canadian Occupational Performance Measure (COPM)							
Functional Independence Measure (FIM)	1		1		1		**3**
Houghton Scale	4	4	4	4	1	1	**18**
Locomotor Capabilities Index (original)	2	1	2	1	1	1	**8**
LCI-5	9	7	11	8	6	5	**46**
Prosthesis Evaluation Questionnaire				1			**1**
Short Form 12 or 36 Health Survey (SF-12 or 36)	1			4			**5**
Special Interest Group in Amputee Medicine (SIGAM) Mobility Grades	11	9	11	8	5	5	**49**
Socket Comfort Score (SCS)	7	7	7	7	5	5	**38**
Timed "Up and Go" Test (TUG)	13	12	12	11	7	5	**60**
Timed Walk Test - 2min or 6min (TWT)	9	10	11	11	6	6	**53**
Trinity Amputation and Prosthesis Experience Scales	2	2	3	5	2	2	**18**

Prosthetists' responses (n=17)

	Activity Level K0 – K2 Non-limb wearer, indoor or limited outdoor ambulator		**Activity Level K3** Wearing limb daily and fully ambulant outdoors		**Activity Level K4** Undertakes athletic activities in addition to daily tasks		**TOTALS**
	Primary	**Established**	**Primary**	**Established**	**Primary**	**Established**	
Activities-specific Balance Confidence (ABC) Scale-UK							
Amputee Activity Score	2	2	2	2	1	2	**11**
Amputee Mobility Predictor with Prosthesis							
Attitude to Artificial Limb Questionnaire	1		1				**2**
Barthel Index							
Canadian Occupational Performance Measure (COPM)							
Functional Independence Measure (FIM)							
Houghton Scale	1		1				**2**
Locomotor Capabilities Index or LCI-5							
Prosthesis Evaluation Questionnaire	1	1	1	1	1	1	**6**
Short Form 12 or 36 Health Survey (SF-12 or 36)							
Special Interest Group in Amputee Medicine (SIGAM) Mobility Grades	5	5	5	5	4	5	**29**
Socket Comfort Score (SCS)	2	3	2	3	2	3	**15**
Timed "Up and Go" Test (TUG)	1		1				**2**
Timed Walk Test - 2min or 6min (TWT)	1	1	1	1			**4**
Trinity Amputation and Prosthesis Experience Scales							

Occupational Therapists' responses (n=5)

	Activity Level K0 – K2 Non-limb wearer, indoor or limited outdoor ambulator		**Activity Level K3** Wearing limb daily and fully ambulant outdoors		**Activity Level K4** Undertakes athletic activities in addition to daily tasks		**TOTALS**
	Primary	**Established**	**Primary**	**Established**	**Primary**	**Established**	
Activities-specific Balance Confidence (ABC) Scale-UK							
Amputee Activity Score							
Amputee Mobility Predictor with Prosthesis							
Attitude to Artificial Limb Questionnaire							
Barthel Index	1						**1**
Canadian Occupational Performance Measure (COPM)	2	1	2				**5**
Functional Independence Measure (FIM)	1		1				**2**
Houghton Scale							
Locomotor Capabilities Index or LCI-5							
Prosthesis Evaluation Questionnaire							
Short Form 12 or 36 Health Survey (SF-12 or 36)							
Special Interest Group in Amputee Medicine (SIGAM) Mobility Grades							
Socket Comfort Score (SCS)							
Timed "Up and Go" Test (TUG)							
Timed Walk Test - 2min or 6min (TWT)							
Trinity Amputation and Prosthesis Experience Scales							

Appendix 4

Filter search terms used

The following search terms were taken from the paper by Terwee et al 2009:

Terwee CB, Jansma EP, Riphagen II, De Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Quality of Life Research 2009;18:1115–23. doi:10.1007/s11136-009-9528-5.

(instrumentation[sh] OR Validation Studies[pt] OR ‘‘reproducibility of results’’[MeSH Terms] OR reproducib*[tiab] OR ‘‘psychometrics’’[MeSH] OR psychometr*[tiab] OR clinimetr*[tiab] OR clinometr*[tiab] OR ‘‘observer variation’’[MeSH] OR observer variation[tiab] OR ‘‘discriminant analysis’’[MeSH] OR reliab*[tiab] OR valid*[tiab] OR coefficient[tiab] OR ‘‘internal consistency’’[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR ‘‘item correlation’’[tiab] OR ‘‘item correlations’’[tiab] OR ‘‘item selection’’[tiab] OR ‘‘item selections’’[tiab] OR ‘‘item reduction’’[tiab] OR ‘‘item reductions’’[tiab] OR agreement[tw] OR precision[tw] OR imprecision[tw] OR ‘‘precise values’’[tw] OR test–retest [tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR intertechnician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[-tiab] OR kappa’s[tiab] OR kappas[tiab] OR ‘‘coefficient of variation’’[tiab] OR repeatab*[tw] OR ((replicab*[tw] OR repeated[tw]) AND (measure[tw] OR measures[tw] OR findings[tw] OR result[tw] OR results[tw] OR test[tw] OR tests[tw])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR ‘‘known group’’ [tiab] OR ‘‘factor analysis’’[tiab] OR ‘‘factor analyses’’[tiab] OR ‘‘factor structure’’[tiab] OR ‘‘factor structures’’[tiab] OR dimensionality[tiab] OR subscale*[tiab] OR ‘‘multitrait scaling analysis’’[tiab] OR ‘‘multitrait scaling analyses’’[
tiab] OR ‘‘item discriminant’’[tiab]OR ‘‘interscale correlation’’[tiab] OR ‘‘interscale correlations’’[tiab] OR ((error[tiab] OR errors[tiab]) AND (measure*[tiab] OR correlat*[tiab] OR evaluat*[tiab] OR accuracy[tiab] OR accurate[tiab] OR precision[tiab] OR mean[tiab])) OR ‘‘individual variability’’[tiab] OR ‘‘interval variability’’[tiab] OR ‘‘rate variability’’[tiab] OR ‘‘variability analysis’’[tiab] OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab])) OR ‘‘standard error of measurement’’[tiab] OR sensitiv*[tiab] OR

responsive*[tiab] OR (limit[tiab] AND detection[tiab]) OR ‘‘minimal detectable concentration’’[tiab]ORinterpretab*[tiab] OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR ‘‘meaningful change’’[tiab] OR ‘‘minimal important change’’[tiab] OR ‘‘minimal important difference’’[tiab] OR ‘‘minimally important change’’[tiab] OR ‘‘minimally important difference’’[tiab] OR ‘‘minimal detectable change’’[tiab] OR ‘‘minimal detectable difference’’[tiab] OR ‘‘minimally detectable change’’[tiab] OR ‘‘minimally detectable difference’’[tiab] OR ‘‘minimal real change’’[tiab] OR ‘‘minimal real difference’’[tiab] OR ‘‘minimally real change’’[tiab] OR ‘‘minimally real difference’’[tiab] OR ‘‘ceiling effect’’[tiab] OR ‘‘floor effect’’ [tiab] OR ‘‘Item response model’’[tiab] OR IRT[tiab] OR Rasch[tiab] OR ‘‘Differential item functioning’’[tiab] OR
DIF[tiab] OR ‘‘computer adaptive testing’’[tiab] OR ‘‘item bank’’[tiab] OR ‘‘cross-cultural equivalence’’[tiab])

Appendix 5

Reviewers’ Notes

Reminder of criteria

Inclusion criteria

- Study population: adult (over 18 years) lower limb amputees, unilateral or bilateral at any level.
- Studies reporting on clinimetric or psychometric properties of OMs of physical function used during any stage of Prosthetic Rehabilitation.

Exclusion criteria

- Dissertations, conference proceedings, editorials, opinion pieces, review papers, letters, single case studies and sample size of five or fewer patients
- Studies that validate translated versions of the outcome measures
- Outcome measures developed or used exclusively (75% rule) for predicting outcomes.
- Outcome measures developed or used exclusively (75% rule) for measuring Quality of Life
- Studies that examine the performance of either prosthetic componentry e.g. different types of knees/feet rather than that of the participants
- Studies that examine the performance of any electronic or digital instrumentation rather than that of the participant, e.g. step counters, accelerometers, gait analysis instrumentation etc.

Discussion points between reviewers with agreed actions

- Missing data:
  If no mention then assume % not described - GOOD.
  If not described then not clear how items were handled - FAIR.

- Sample size:
  Given that the sample sizes can be small because of the relatively small numbers of lower limb amputees will consider reporting results with and without taking the sample size into account.

- Time Interval: (boxes B & C)
  Consider the population under study and the parameter being studied when considering whether appropriate or justified.

- Patient's stability: (boxes B & C)
  Evidence should be provided in the text to be graded EXCELLENT.

- Independent administration: (boxes B & C)
  Evidence should be provided that steps were taken for the tester or the subject to have no knowledge of previous results and/or there was appropriate time interval (see above) to be graded EXCELLENT.

- ICC calculations: (boxes B & C)
  The model/formula should have been mentioned in the text to achieve an EXCELLENT grading.

- Comparator Instrument: (boxes F & I – where no Gold Standard)
  - Description of the instrument in the study text must have references pertaining to the study population to achieve an EXCELLENT grading. If the title of the reference does not make it clear re: study population then the grading should be GOOD. If unsure then it should be FAIR or even POOR. NB some references may be included in the review and therefore even if the title is not clear the grading may be higher as a consequence. These instances will be noted.
  - Similar for evidence of the measurement properties of the comparator.

Appendix 6

COSMIN 4-point checklist

The checklist included in this appendix is available to download from the COSMIN website http://www.cosmin.nl/Publications.html and is described in detail in the following article:

Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of Life Research 2012;21:651-657

COSMIN checklist with 4-point scale

Contact
CB Terwee, PhD
VU University Medical Center
Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
1081 BT Amsterdam
The Netherlands
Website: www.cosmin.nl, www.emgo.nl
E-mail: cb.terwee@vumc.nl

Instructions

This version of the COSMIN checklist is recommended for use in systematic reviews of measurement properties. With this version it is possible to calculate overall methodological quality scores per study on a measurement property. A methodological quality score per box is obtained by taking the lowest rating of any item in a box ('worse score counts'). For example, if for a reliability study one item in the box 'Reliability' is scored poor, the methodological quality of that reliability study is rated as poor. The Interpretability box and the Generalizability box are mainly used as data extraction forms. We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box (e.g. norm scores, floor-ceiling effects, minimal important change) of the instruments under study from the included articles. Similar, we recommend to use the Generalizability box to extract data on the characteristics of the study population and sampling procedure. Therefore no scoring system was developed for these boxes.

This scoring system is described in this paper:

Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of Life Research 2011, July 6 [epub ahead of print].

Step 1. Evaluated measurement properties in the article

	Internal consistency	Box A
	Reliability	Box B
	Measurement error	Box C
	Content validity	Box D
	Structural validity	Box E
	Hypotheses testing	Box F
	Cross-cultural validity	Box G
	Criterion validity	Box H
	Responsiveness	Box I

Step 2. Determining if the statistical method used in the article are based on CTT or IRT

Box General requirements for studies that applied Item Response Theory (IRT) models					
		excellent	**good**	**fair**	**poor**
1	Was the IRT model used adequately described? e.g. One Parameter Logistic Model (OPLM), Partial Credit Model (PCM), Graded Response Model (GRM)	IRT model adequately described	IRT model not adequately described		
2	Was the computer software package used adequately described? e.g. RUMM2020, WINSTEPS, OPLM, MULTILOG, PARSCALE, BILOG, NLMIXED	Software package adequately described	Software package not adequately described		
3	Was the method of estimation used adequately described? e.g. conditional maximum likelihood (CML), marginal maximum likelihood (MML)	Method of estimation adequately described	Method of estimation not adequately described		
4	Were the assumptions for estimating parameters of the IRT model checked? e.g. unidimensionality, local independence, and item fit (e.g. differential item functioning (DIF))	assumptions of the IRT model checked	assumptions of the IRT model partly checked	assumptions of the IRT model not checked or unknown	

To obtain a total score for the methodological quality of studies that use IRT methods, the ‘worse score counts’ algorithm should be applied to the IRT box in combination with the box of the measurement property that was evaluated in the IRT study. For example, if IRT methods are used to study internal consistency and item 4 in the IRT box is scored fair, while the items in the internal consistency box (box A) are all scored as good or excellent, the methodological quality score for internal consistency will be fair. However, if any of the items in box A is scored poor, the methodological quality score for internal consistency will be poor.

Step 3. Determining if a study meets the standards for good methodological quality

Box A. Internal consistency				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?				
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items

7	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
9	for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10	for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11	for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate

9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box C. Measurement error: absolute measures				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate

9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for CTT: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	SEM, SDC, or LoA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SD from another population

Box D. Content validity (including face validity)					
		excellent	**good**	**fair**	**poor**
General requirements					
1	Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured

2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box E. Structural validity				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?				
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 5-7* #items but <100	5* #items but <100	<5* #items
5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. rotation method not described)	Other important methodological flaws in the design or execution of the study (e.g. inappropriate rotation method)

Statistical methods					
6	for CTT: Was exploratory or confirmatory factor analysis performed?	Exploratory or confirmatory factor analysis performed and type of factor analysis appropriate in view of existing information	Exploratory factor analysis performed while confirmatory would have been more appropriate		No exploratory or confirmatory factor analysis performed
7	for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	IRT test for determining (uni)dimension-ality performed			IRT test for determining (uni)dimension-ality NOT performed

Box F. Hypotheses testing					
		excellent	**good**	**fair**	**Poor**
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)

4	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5	Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6	Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)

9 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods				
10 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box G. Cross-cultural validity				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	

3	Was the sample size included in the analysis adequate?	CTT: 7* #items and ≥100 IRT: ≥200 per group	CTT: 5* #items and ≥100 OR 5-7* #items but <100 IRT: ≥200 in 1 group and 100-199 in 1 group	CTT: 5* #items but <100 IRT: 100-199 per group	CTT: <5* #items IRT: (<100 in 1 or both groups
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	Both source language and target language described			Source language NOT known
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	Expertise of the translators described with respect to disease, construct, and language	Expertise of the translators with respect to disease or construct poor or not described	Expertise of the translators with respect to language not described	
6	Did the translators work independently from each other?	Translators worked independent	Assumable that the translators worked independent	Unclear whether translators worked independent	Translators worked NOT independent
7	Were items translated forward and backward?	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation
8	Was there an adequate description of how differences between the original and translated versions were resolved?	Adequate description of how differences between translators were resolved	Poorly or NOT described how differences between translators were resolved		

9	Was the translation reviewed by a committee (e.g. original developers)?	Translation reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation NOT reviewed by (such) a committee		
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	Translated instrument pre-tested in the target population	Translated instrument pre-tested, but unclear if this was done in the target population	Translated instrument pre-tested, but NOT in the target population	Translated instrument NOT pre-tested
11	Was the sample used in the pre-test adequately described?	Sample used in the pre-test adequately described		Sample used in the pre-test NOT (adequately) described	
12	Were the samples similar for all characteristics except language and/or cultural background?	Shown that samples were similar for all characteristics except language /culture	Stated (but not shown) that samples were similar for all characteristics except language /culture	Unclear whether samples were similar for all characteristics except language /culture	Samples were NOT similar for all characteristics except language /culture
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Statistical methods		
14 for CTT: Was confirmatory factor analysis performed?	Multiple-group confirmatory factor analysis performed	Multiple-group confirmatory factor analysis NOT performed
15 for IRT: Was differential item function (DIF) between language groups assessed?	DIF between language groups assessed	DIF between language groups NOT assessed

Box H. Criterion validity

	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'

5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
6	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated		Correlations or AUC NOT calculated
7	for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated		Sensitivity and specificity NOT calculated

Box I. Responsiveness				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
5 Was the time interval stated?	Time interval adequately described			Time interval NOT described

6	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7	Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
Design requirements for hypotheses testing					
	For constructs for which a gold standard was not available:				
8	Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
9	Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10	Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		

11	Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
12	Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods					
14	Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Design requirement for comparison to a gold standard				
	For constructs for which a gold standard was available:			
15	Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
16	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
17	for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Area under the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated
18	for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Interpretability

We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box of the instruments under study from the included articles.

Box Interpretability	
Percentage of missing items	
Description of how missing items were handled	
Distribution of the (total) scores	
Percentage of the respondents who had the lowest possible (total) score	
Percentage of the respondents who had the highest possible (total) score	
Scores and change scores (i.e. means and SD) for relevant (sub) groups, e.g. for normative groups, subgroups of patients, or the general population	
Minimal Important Change (MIC) or Minimal Important Difference (MID)	

Generalizability

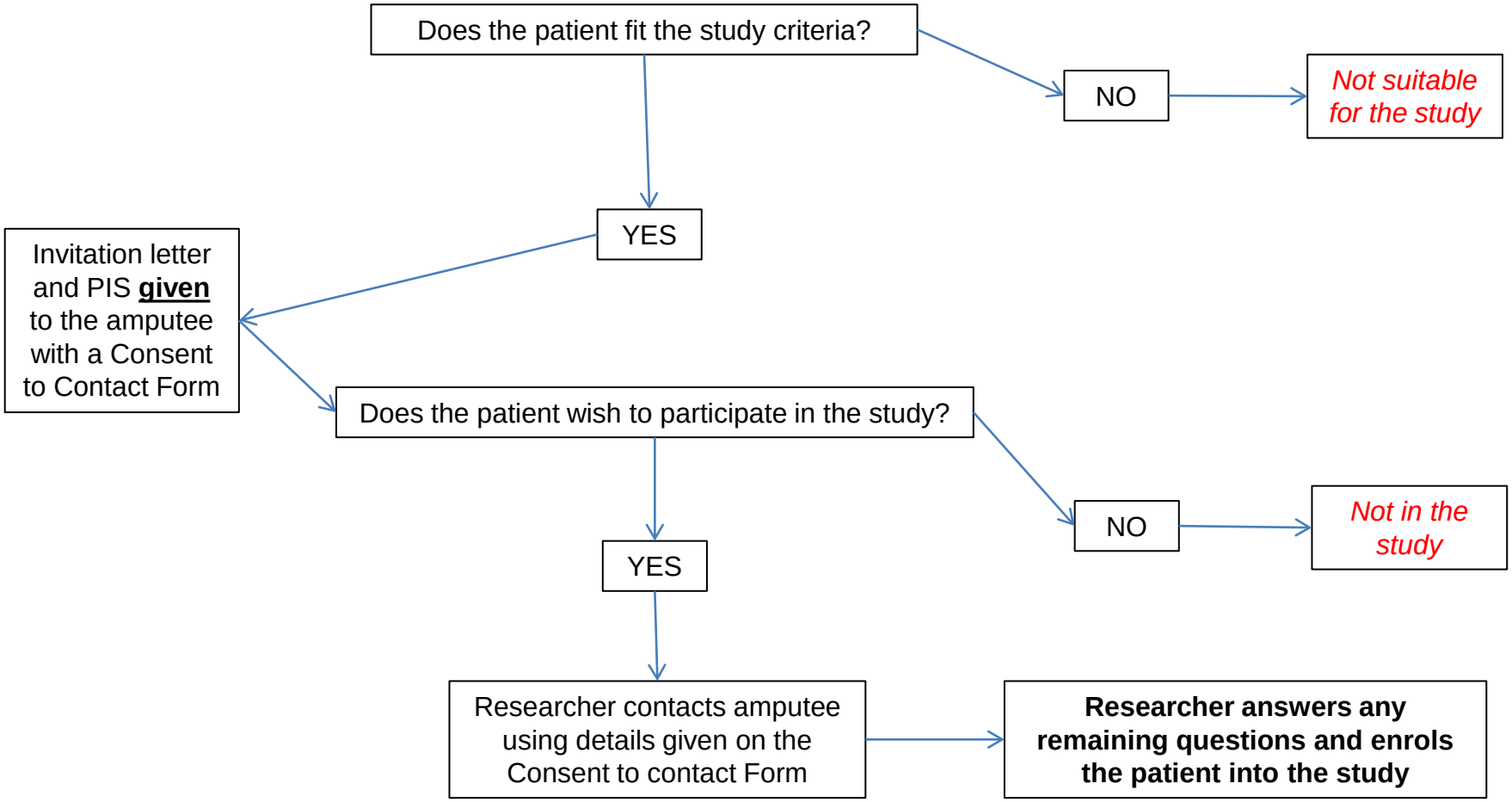
We recommend to use the Generalizability box to extract data on the characteristics of the study populations and sampling procedures of the included studies.

Box Generalisability	
Median or mean age (with standard deviation or range)	
Distribution of sex	
Important disease characteristics (e.g. severity, status, duration) and description of treatment	
Setting(s) in which the study was conducted (e.g. general population, primary care or hospital/rehabilitation care)	
Countries in which the study was conducted	
Language in which the HR-PRO instrument was evaluated	
Method used to select patients (e.g. convenience, consecutive, or random)	
Percentage of missing responses (response rate)	

Study Title: Outcome Measures for Amputees - A Repeatability Study

Queen Margaret University
EDINBURGH

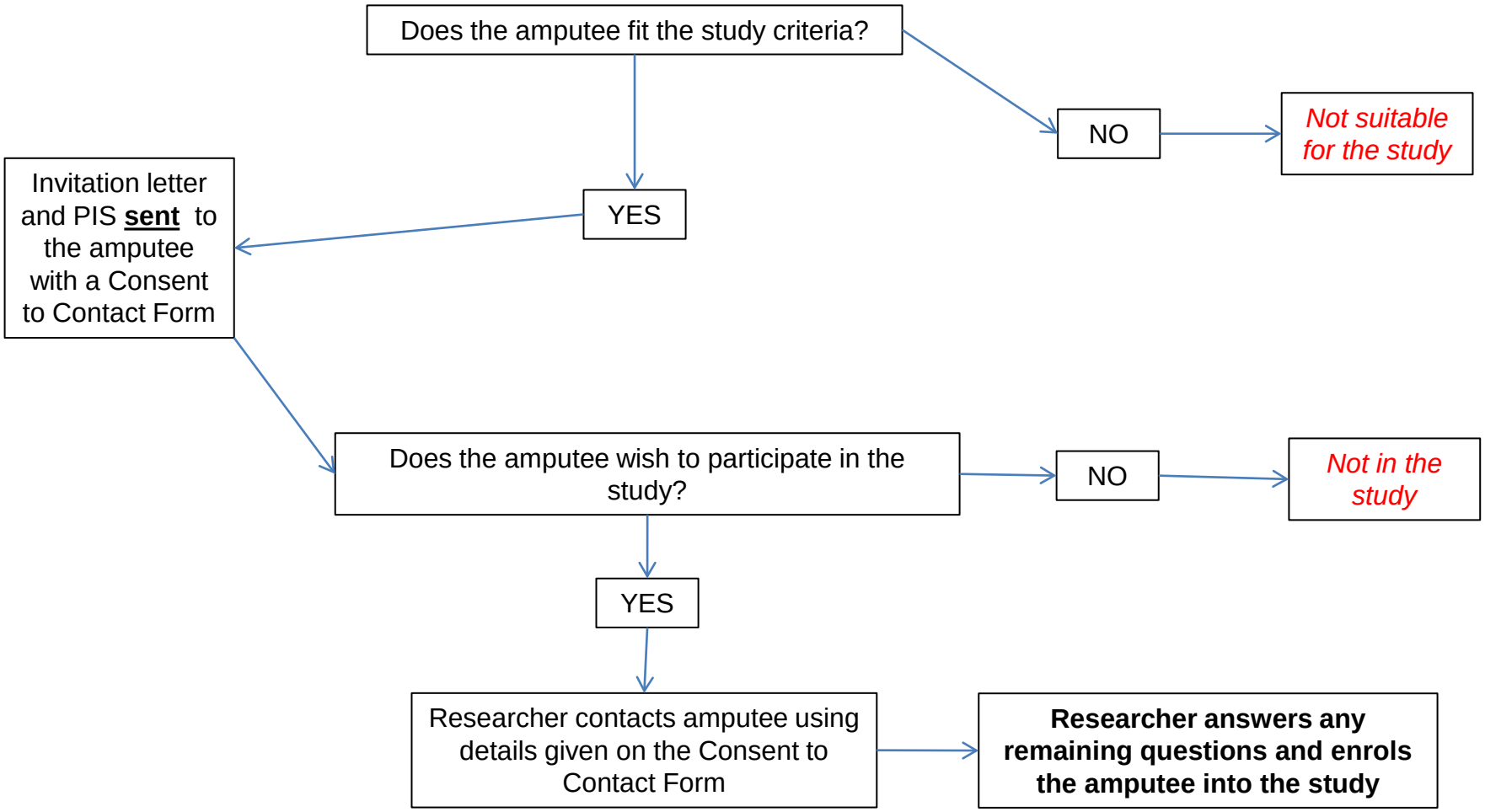
Recruitment Decision Process from Advert



Study Title: Outcome Measures for Amputees - A Repeatability Study

Queen Margaret University
EDINBURGH

Recruitment Decision Process from Database Search



Appendix 8

Recruitment adverts – study III

Queen Margaret University
EDINBURGH

Would you like to take part in a study looking into Activity Outcome Measures?

If you have had your artificial leg for over a year,

you wear it all day, 

and

you are active outdoors, 

you may be eligible to help with this study.

The study involves visiting the Physiotherapy Department here at Astley Ainslie Hospital on 2 separate occasions, a week apart.

At each visit you will be asked to:

- **complete 4 questionnaires about your activity levels and**
- **participate in two walking tasks.**

The visits should take no longer than 1 hour of your time and you will receive a £20 voucher when you complete the second visit.

If you are interested in hearing more please contact your Prosthetist for an Information Study Pack.

THANK YOU

Judy Scopes, PhD Student, Queen Margaret University, Edinburgh

The study is part of a Physiotherapy PhD taking place at Queen Margaret University, Edinburgh.

Queen Margaret University
EDINBURGH

Would you like to take part in a study looking into Activity Outcome Measures?

If you have had your artificial leg for over a year,

you wear it all day, 

and

you are active outdoors, 

you may be eligible to help with this study.

The study involves visiting the Physiotherapy Department here at Astley Ainslie Hospital on 2 separate occasions, a week apart.

At each visit you will be asked to:

- **complete 4 questionnaires about your activity levels and**
- **participate in two walking tasks.**

The visits should take no longer than 1 hour of your time and you will receive a £20 voucher when you complete the second visit.

If you are interested in hearing more please contact the Researcher, Judy Scopes on JScopes@qmu.ac.uk for an Information Study Pack.

THANK YOU

Judy Scopes, PhD Student, Queen Margaret University, Edinburgh

The study is part of a Physiotherapy PhD taking place at Queen Margaret University, Edinburgh.

Study Title: Outcome Measures for Amputees - A Repeatability Study

Queen Margaret University
EDINBURGH

Study Schedule

Test Visit 1 (TV1)

Written Consent obtained

Data Collected at TV1:

Demographic Data

Outcome Measures:
SIGAM
LCI-5
TUAG
2minTWT
SCS
EQ-5D

Test Visit 2 (TV2)

Within 7-10 days of visit 1

Data Collected at TV2:

Outcome Measures:
SIGAM
LCI-5
TUAG
2minTWT
SCS
EQ-5D

Study Title: Outcome Measures for Amputees - A Repeatability Study

Queen Margaret University
EDINBURGH

Participant No:

Individual Participant Study Schedule

Test Visit 1 (SV1)
Date:
Venue:
Written Consent
Demographic Data
SIGAM
LCI-5
TUAG
2minTWT
SCS
EQ-5D

Test Visit 2 (SV2)
Date:
Venue:
SIGAM
LCI-5
TUAG
2minTWT
SCS
EQ-5D

Appendix 10

Data collection sheets and instructions for outcome measures

Study III - Repeatability Study

Study IV – Longitudinal Cohort Study (Rehabilitation Study)

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Repeatability Study

Demographic Data

Participant Number:	

Age		
Cause of Amputation	PVD	
	Diabetes	
	Trauma	
	Tumour	
	Other	
Level of Amputation	TTA	
	TFA	
Prosthetic usage	Time since delivery of first artificial leg	
Other relevant Medical History	PVD	
	Diabetes	
	Heart Disease	
	Vascular Disease	
	Blood Pressure	
	Neurological conditions	
	Arthritis	
	Liver disease	

GP Name: _____

GP Address: _____

_____

_____

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Repeatability Study

SIGAM Mobility Grades

Participant No:		Date:		Test Visit:	

SIGAM Mobility Grade Recorded	

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Repeatabilty Study

SIGAM mobility grades Algorithm

Guidelines for the use of SIGAM Algorithm

1. Start at the top of page at question 1. The number beside each box corresponds to the same numbered question.
2. Depending on the answer "YES" or "NO" follow the arrows to assign one of Grade A – E.
3. Once you reach | Grade….. | then that is the grade to be assigned. You do not need to proceed, except for Grades C & D to assign a sub-grade.
4. Sub-grades (a) – (d) apply depending on what walking aid is used to assist walking. If a patient ticks "Yes" to more than one aid then he/she is graded based upon that which provides most support: frame > 2 crutches/sticks > 1 crutch/stick > none

Reproduced from: RYALL, N.H., et al, 2003. The SIGAM Mobility Grades: A New Population-Specific Measure for Lower Limb Amputees. *Disability and Rehabilitation*, vol. 25, no. 15, pp. 833-844.

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Repeatability Study

LOCOMOTOR CAPABILITIES INDEX - 5

Whether or not you wear your prosthesis, at the present time, would you say that you are "able" to do the following activities WITH YOUR PROSTHESIS ON?

Please **circle the number** that best describes your capability.

ITEM	NO	YES, if someone helps me	YES, if someone is near me	YES, alone, with ambulation aids	YES, alone, without ambulation aids
1. Get up from a chair	0	1	2	3	4
2. Walk in the house	0	1	2	3	4
3. Walk outside on even ground	0	1	2	3	4
4. Go up the stairs with a handrail	0	1	2	3	4
5. Go down the stairs with a handrail	0	1	2	3	4
6. Step up a sidewalk curb	0	1	2	3	4
7. Step down a sidewalk curb	0	1	2	3	4
Basic Activities Score					
1. Pick up an object from the floor (when you are standing up with your prosthesis)	0	1	2	3	4
2. Get up from the floor (e.g. if you fall)	0	1	2	3	4
3. Walk outside on uneven ground (e.g. grass, gravel, slope)	0	1	2	3	4
4. Walk outside in inclement weather (e.g. snow, rain, ice)	0	1	2	3	4
5. Go up a few steps (stairs) without a handrail	0	1	2	3	4
6. Go down a few steps (stairs) without a handrail	0	1	2	3	4
7. Walk while carrying an object.	0	1	2	3	4
Advanced Activities Score					
Total Score					

Participant No:		Date:		Test Visit:	

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Repeatability Study

Timed up and Go Test

Test Protocol

1. **Test setup**
   An upright armchair is positioned in the gym area with a cone placed on the floor 3m away directly in front of the chair.
2. **Subject preparation**
   The participant is asked to sit in the chair and position themselves with any walking aids ready. The position of any prosthetic knee joints will be noted as either locked or unlocked prior to starting the test and the same position will be taken on any subsequent test.
3. **Explanation of the test**
   The following instructions will be given: "Sit with your back against the chair and your arms on the armrest. When I say GO please get up and using your walking aid, walk around the cone, then return to the chair and sit back down. I am going to be timing, but this is not a race, please go at a pace that's comfortable and safe for you."
4. **Timing**
   Timing will start on the word GO and will stop when the participant's buttocks first touch the seat.
5. **Repetitions**
   If the participant is able, the test is repeated 3 times, giving time to recover between attempts. The first attempt is used for familiarisation and checking they have understood the instructions. The second and third times will be recorded and the faster of the two will be the score. Note if only one attempt is possible and this will be the score.

Participant No:		Date:		Test Visit:	

Attempt No:	Knee position	Walking aid used	Time (secs)
1			
2			
3			

Recorded Score (secs)	

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Repeatability Study

2min Timed Walk Test

Test Protocol

1. **Test setup**
   A hallway free of obstacles with pre-measured distances marked for easy calculation of the total distance covered.

2. **Subject preparation**
   The position and type of any prosthetic knee joints will be noted at every test. The type of walking aid used will also be noted as applicable.

3. **Explanation of the test**
   The following instructions will be given to the participant: "*Cover as much ground as possible over 2 minutes. Walk continuously if possible, but do not be concerned if you need to slow down or stop to rest. The goal is to feel at the end of the test that more ground could not have been covered in the 2 minutes."*

4. **Timing**
   Timing will start when the participant starts walking and the distance covered will be measured at the end of 2 minutes..

Participant No:		Date:		Test Visit:	

Knee position	Walking aid used	Time (secs)

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Repeatability Study

Socket Comfort Score

Participant No:		Date:		Test Visit:	

On a scale of 0 – 10,
if 0 represents the most uncomfortable socket fit you can imagine,
and 10 represents the most comfortable socket fit,
how would you score the comfort of the socket fit of your artificial limb
at the moment

0
1
2
3
4
5
6
7
8
9
10

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

Demographic Data

Participant Number:	

Age		
Cause of Amputation	PVD	
	Diabetes	
	Trauma	
	Tumour	
	Other	
Level of Amputation	TTA	
	TFA	
Prosthetic usage	Time since delivery of first artificial leg	
Other relevant Medical History	PVD	
	Diabetes	
	Heart Disease	
	Vascular Disease	
	Blood Pressure	
	Neurological conditions	
	Arthritis	
	Liver disease	

GP Name: _____

GP Address: _____

_____

_____

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

Activity Change Questionnaire SV2

Participant No:		Date:		Study Visit:	2

Please indicate, by circling one of the statements below, how you feel your current physical ability to perform everyday tasks is today, compared to when you first got your artificial leg.

Much worse	Slightly worse	The same	Slightly better	Much better

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

Activity Change Questionnaire SV4a

Participant No:		Date:		Study Visit:	4

Please indicate, by circling one of the statements below, how you feel your current physical ability to perform everyday tasks is today, compared to when you first got your artificial leg

Much worse	Slightly worse	The same	Slightly better	Much better

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

Activity Change Questionnaire SV4b

Participant No:		Date:		Study Visit:	4

Please indicate, by circling one of the statements below, how you feel your current physical ability to perform everyday tasks is today, compared to when you left hospital.

Much worse	Slightly worse	The same	Slightly better	Much better

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

SIGAM Mobility Grades

Participant No:		Date:		Test Visit:	

SIGAM Mobility Grade Recorded	

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

SIGAM mobility grades Algorithm

Guidelines for the use of SIGAM Algorithm

5. Start at the top of page at question 1. The number beside each box corresponds to the same numbered question.
6. Depending on the answer "YES" or "NO" follow the arrows to assign one of Grade A – E.
7. Once you reach | Grade….. | then that is the grade to be assigned. You do not need to proceed, except for Grades C & D to assign a sub-grade.
8. Sub-grades (a) – (d) apply depending on what walking aid is used to assist walking. If a patient ticks "Yes" to more than one aid then he/she is graded based upon that which provides most support: frame > 2 crutches/sticks > 1 crutch/stick > none

Reproduced from: RYALL, N.H., et al, 2003. The SIGAM Mobility Grades: A New Population-Specific Measure for Lower Limb Amputees. *Disability and Rehabilitation*, vol. 25, no. 15, pp. 833-844.

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

LOCOMOTOR CAPABILITIES INDEX - 5

Whether or not you wear your prosthesis, at the present time, would you say that you are "able" to do the following activities WITH YOUR PROSTHESIS ON?

Please **circle the number** that best describes your capability.

ITEM	NO	YES, if someone helps me	YES, if someone is near me	YES, alone, with ambulation aids	YES, alone, without ambulation aids
1. Get up from a chair	0	1	2	3	4
2. Walk in the house	0	1	2	3	4
3. Walk outside on even ground	0	1	2	3	4
4. Go up the stairs with a handrail	0	1	2	3	4
5. Go down the stairs with a handrail	0	1	2	3	4
6. Step up a sidewalk curb	0	1	2	3	4
7. Step down a sidewalk curb	0	1	2	3	4
Basic Activities Score					
1. Pick up an object from the floor (when you are standing up with your prosthesis)	0	1	2	3	4
2. Get up from the floor (e.g. if you fall)	0	1	2	3	4
3. Walk outside on uneven ground (e.g. grass, gravel, slope)	0	1	2	3	4
4. Walk outside in inclement weather (e.g. snow, rain, ice)	0	1	2	3	4
5. Go up a few steps (stairs) without a handrail	0	1	2	3	4
6. Go down a few steps (stairs) without a handrail	0	1	2	3	4
7. Walk while carrying an object.	0	1	2	3	4
Advanced Activities Score					
Total Score					

Participant No:		Date:		Test Visit:	

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

Timed up and Go Test

Test Protocol

6. **Test setup**
   An upright armchair is positioned in the gym area with a cone placed on the floor 3m away directly in front of the chair.
7. **Subject preparation**
   The participant is asked to sit in the chair and position themselves with any walking aids ready. The position of any prosthetic knee joints will be noted as either locked or unlocked prior to starting the test and the same position will be taken on any subsequent test.
8. **Explanation of the test**
   The following instructions will be given: "Sit with your back against the chair and your arms on the armrest. When I say GO please get up and using your walking aid, walk around the cone, then return to the chair and sit back down. I am going to be timing, but this is not a race, please go at a pace that's comfortable and safe for you."
9. **Timing**
   Timing will start on the word GO and will stop when the participant's buttocks first touch the seat.
10. **Repetitions**
    If the participant is able, the test is repeated 3 times, giving time to recover between attempts. The first attempt is used for familiarisation and checking they have understood the instructions. The second and third times will be recorded and the faster of the two will be the score. Note if only one attempt is possible and this will be the score.

Participant No:		Date:		Test Visit:	

Attempt No:	Knee position	Walking aid used	Time (secs)
1			
2			
3			

Recorded Score (secs)	

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

2min Timed Walk Test

Test Protocol

5. **Test setup**
   A hallway free of obstacles with pre-measured distances marked for easy calculation of the total distance covered.

6. **Subject preparation**
   The position and type of any prosthetic knee joints will be noted at every test. The type of walking aid used will also be noted as applicable.

7. **Explanation of the test**
   The following instructions will be given to the participant: "*Cover as much ground as possible over 2 minutes. Walk continuously if possible, but do not be concerned if you need to slow down or stop to rest. The goal is to feel at the end of the test that more ground could not have been covered in the 2 minutes."*

8. **Timing**
   Timing will start when the participant starts walking and the distance covered will be measured at the end of 2 minutes..

Participant No:		Date:		Test Visit:	

Knee position	Walking aid used	Time (secs)

Queen Margaret University
EDINBURGH

Study Title: Outcome Measures for Amputees - A Rehabilitation Study

Socket Comfort Score

Participant No:		Date:		Test Visit:	

On a scale of 0 – 10,
if 0 represents the most uncomfortable socket fit you can imagine,
and 10 represents the most comfortable socket fit,
how would you score the comfort of the socket fit of your artificial limb
at the moment

0
1
2
3
4
5
6
7
8
9
10

Health Questionnaire

English version for the UK

Under each heading, please tick the ONE box that best describes your health TODAY

MOBILITY

I have no problems in walking about ❑
I have slight problems in walking about ❑
I have moderate problems in walking about ❑
I have severe problems in walking about ❑
I am unable to walk about ❑

SELF-CARE

I have no problems washing or dressing myself ❑
I have slight problems washing or dressing myself ❑
I have moderate problems washing or dressing myself ❑
I have severe problems washing or dressing myself ❑
I am unable to wash or dress myself ❑

USUAL ACTIVITIES *(e.g. work, study, housework, family or leisure activities)*

I have no problems doing my usual activities ❑
I have slight problems doing my usual activities ❑
I have moderate problems doing my usual activities ❑
I have severe problems doing my usual activities ❑
I am unable to do my usual activities ❑

PAIN / DISCOMFORT

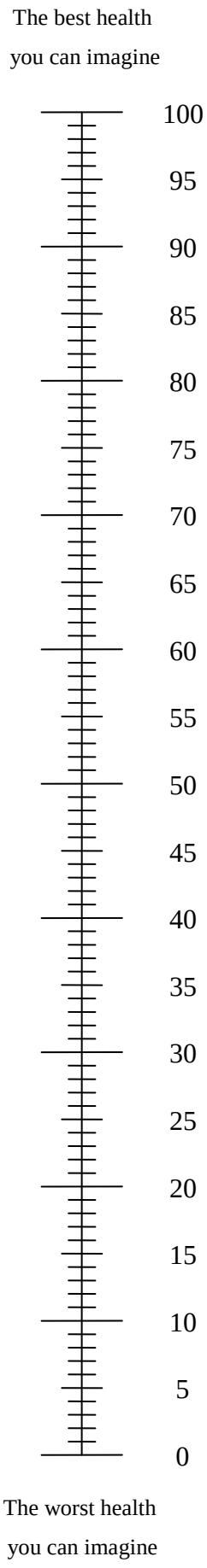
I have no pain or discomfort ❑
I have slight pain or discomfort ❑
I have moderate pain or discomfort ❑
I have severe pain or discomfort ❑
I have extreme pain or discomfort ❑

ANXIETY / DEPRESSION

I am not anxious or depressed ❑
I am slightly anxious or depressed ❑
I am moderately anxious or depressed ❑
I am severely anxious or depressed ❑

- We would like to know how good or bad your health is TODAY.
- This scale is numbered from 0 to 100.
- 100 means the best health you can imagine.
  0 means the worst health you can imagine.
- Mark an X on the scale to indicate how your health is TODAY.
- Now, please write the number you marked on the scale in the box below.

YOUR HEALTH TODAY = ☐



Appendix 11

Normality testing results from SPSS

Study III – Repeatability Study

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SIGAMtv1	.254	12	.032	.851	12	.038
SIGAMtv2	.199	12	.200*	.872	12	.068
LCI5tv1	.259	12	.025	.782	12	.006
LCI5tv2	.267	12	.018	.779	12	.005
TUGtv1	.362	12	.000	.585	12	.000
TUGtv2	.388	12	.000	.554	12	.000
Walktesttv1	.199	12	.200*	.949	12	.617
Walktesttv2	.206	12	.172	.939	12	.480
SCStv1	.261	12	.024	.869	12	.063
SCStv2	.262	12	.022	.876	12	.077
EQ5Dindextv1	.262	12	.022	.825	12	.018
EQ5Dindextv2	.276	12	.012	.795	12	.008
EQ5DVAStv1	.271	12	.015	.844	12	.031
EQ5DVAStv2	.250	12	.038	.758	12	.003

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Appendix 12

Test visit data

Study III – Repeatability Study

Test Visit 1

	Visit Date	SIGAM Mobility Grade	Assigned scores	LCI-5	TUAG	2MTWT	SCS	EQ-5D index	assigned score	EQ-5D VAS
P01R	23/10/2013	E	10	54	11.42	111.70	8	21121	0.767	95
P02R	31/10/2013	E	10	55	10.17	118.75	9	22222	0.592	80
P03R	28/11/2013	E	10	52	7.47	150.30	9	11111	1.000	95
P04R	30/01/2014	Dc	9	50	8.84	122.1	9	22223	0.577	70
P05R	05/03/2014	E	10	55	9.62	119.1	10	21221	0.735	100
P06R	03/07/2014	F	11	56	10.1	140	10	11111	1.000	80
P07R	09/07/2014	E	10	52	10.27	112.4	7	11121	0.837	80
P08R	13/08/2014	Dc	9	33	15.16	88.6	7	23222	0.575	80
P09R	27/08/2014	F	11	56	9.78	148.2	8	11111	1.000	80
P10R	01/10/2014	F	11	56	7.36	161.9	7	11111	1.000	95
P11R	14/10/2014	Da	7	39	32.05	65.6	9	21221	0.735	95
P12R	21/10/2014	Dc	9	45	10.55	112.9	9	11111	1.000	100

Test Visit 2

	Visit Date	SIGAM Mobility Grade	Assigned scores	LCI-5	TUAG	2MTWT	SCS	EQ-5D index	assigned score	EQ-5D VAS
P01R	30/10/2013	Dc	9	55	10.55	113.45	9	21121	0.767	95
P02R	07/11/2013	E	10	55	9.27	129.60	9	11111	1.000	80
P03R	05/12/2013	E	10	53	7.95	142.65	9	11111	1.000	95
P04R	06/02/2014	Dc	9	47	7.87	135.7	5	22342	0.348	35
P05R	12/03/2014	E	10	52	10.23	105.8	10	21231	0.710	100
P06R	10/07/2014	F	11	56	10.47	140	10	11111	1.000	70
P07R	16/07/2014	E	10	52	10.64	126.5	7	11121	0.837	90
P08R	20/08/2014	Dc	9	30	13.59	89	8	22222	0.592	80
P09R	03/09/2014	F	11	56	9.16	143.8	9	11111	1.000	80
P10R	08/10/2014	F	11	56	7.96	160.7	8	11111	1.000	95
P11R	21/10/2014	Da	7	41	30.24	72.65	7	12111	0.846	95
P12R	27/10/2014	Dc	9	47	9.91	128.9	9	11111	1.000	90

Appendix 13

Normality testing results from SPSS

Study IV – Longitudinal Cohort Study

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SIGAMsv1	.458	23	.000	.618	23	.000
SIGAMsv2	.369	23	.000	.728	23	.000
SIGAMsv3	.319	23	.000	.803	23	.000

a. Lilliefors Significance Correction

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
LCIsv1	.169	23	.085	.894	23	.019
LCIsv2	.159	23	.136	.919	23	.065
LCIsv3	.140	23	.200*	.904	23	.031

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
TUGsv1	.152	22	.200*	.940	22	.197
TUGsv2	.243	22	.002	.870	22	.008
TUGsv3	.155	22	.186	.901	22	.032

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Walktestsv1	.144	22	.200*	.940	22	.198
Walktestsv2	.181	22	.059	.911	22	.049
Walktestsv3	.121	22	.200*	.940	22	.195

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SCSsv1	.259	23	.000	.906	23	.034
SCSsv2	.270	23	.000	.865	23	.005
SCSsv3	.250	23	.001	.892	23	.018

a. Lilliefors Significance Correction

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
EQ5Dindexsv1	.112	23	.200*	.978	23	.864
EQ5Dindexsv2	.109	23	.200*	.969	23	.670
EQ5Dindexsv3	.139	23	.200*	.925	23	.087

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
EQ5DVASsv1	.159	23	.135	.940	23	.184
EQ5DVASsv2	.184	23	.043	.868	23	.006
EQ5DVASsv3	.136	23	.200*	.961	23	.490

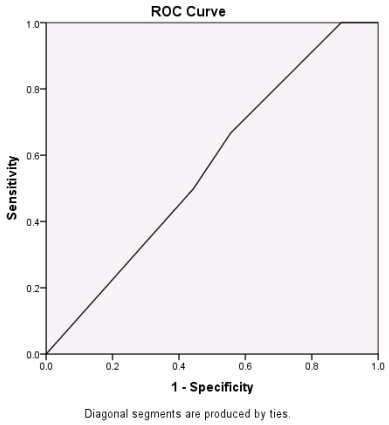
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Appendix 14

ROC curve graphs from SPSS

SIGAM TI1



Area Under the Curve
.565

Coordinates of the Curve

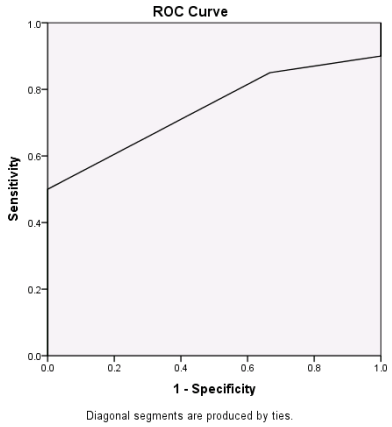
Test Result Variable(s): SIGAMdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-5.0000	1.000	1.000
-2.0000	1.000	.889
.5000	.667	.556
2.5000	.500	.444
5.0000	.000	.000

The test result variable(s): SIGAMdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

SIGAM TI2



Area Under the Curve
.742

Coordinates of the Curve

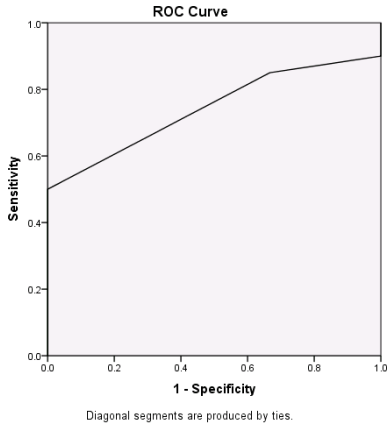
Test Result Variable(s): SIGAMdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-5.0000	1.000	1.000
-3.5000	.950	1.000
-2.5000	.900	1.000
-1.0000	.850	.667
.5000	.500	.000
1.5000	.300	.000
3.0000	.250	.000
4.5000	.150	.000
6.0000	.000	.000

The test result variable(s): SIGAMdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

SIGAM TI3



Area Under the Curve
.546

Coordinates of the Curve

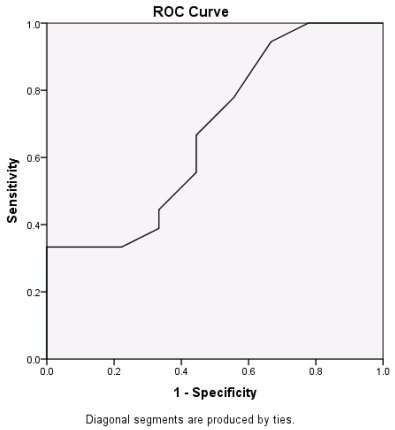
Test Result Variable(s): SIGAMdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-4.0000	1.000	1.000
-2.5000	1.000	.889
-1.0000	1.000	.778
.5000	.882	.778
1.5000	.647	.444
3.0000	.529	.444
4.5000	.176	.333
6.0000	.000	.000

The test result variable(s): SIGAMdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

LCI-5 TI1



Area Under the Curve
.676

Coordinates of the Curve

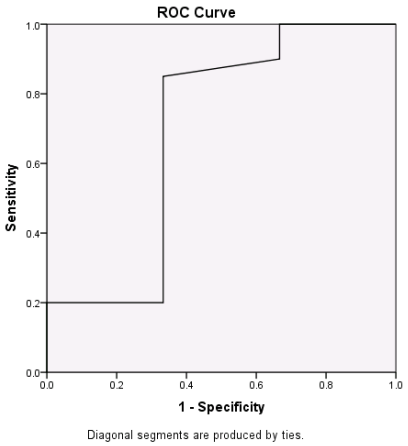
Test Result Variable(s): LCI5diff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-2.0000	1.000	1.000
-.5000	1.000	.778
.5000	.944	.667
1.5000	.778	.556
2.5000	.667	.444
3.5000	.556	.444
4.5000	.444	.333
5.5000	.389	.333
7.0000	.333	.222
8.5000	.333	.111
9.5000	.333	.000
11.5000	.278	.000
16.0000	.222	.000
20.0000	.167	.000
21.5000	.111	.000
23.0000	.056	.000
25.0000	.000	.000

The test result variable(s): LCI5diff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

LCI-5 TI2



Area Under the Curve
.692

Coordinates of the Curve

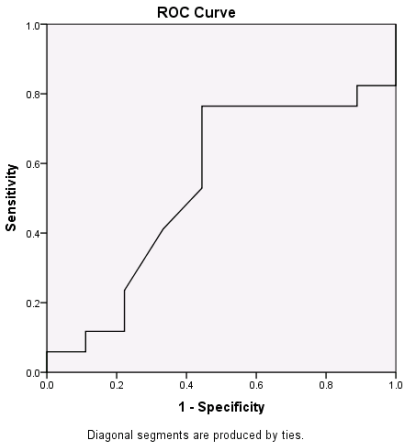
Test Result Variable(s): LCI5diff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-8.0000	1.000	1.000
-6.5000	1.000	.667
-4.5000	.950	.667
-2.5000	.900	.667
-1.5000	.850	.333
.0000	.700	.333
1.5000	.600	.333
2.5000	.500	.333
3.5000	.450	.333
4.5000	.400	.333
6.5000	.300	.333
9.0000	.250	.333
11.0000	.200	.333
12.5000	.200	.000
13.5000	.150	.000
14.5000	.100	.000
17.0000	.050	.000
20.0000	.000	.000

The test result variable(s): LCI5diff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

LCI-5 TI3



Area Under the Curve
.539

Coordinates of the Curve

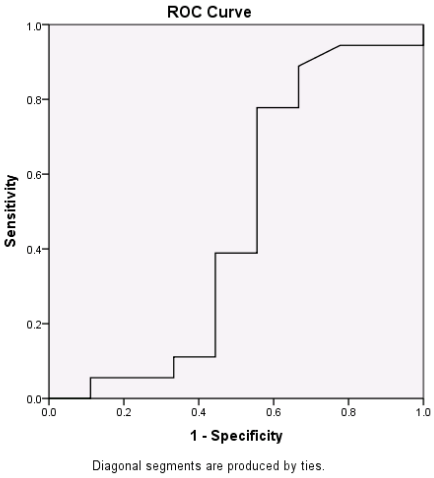
Test Result Variable(s): LCI5diff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
.0000	1.000	1.000
1.5000	.824	1.000
2.5000	.824	.889
3.5000	.765	.889
4.5000	.765	.667
6.0000	.765	.444
8.5000	.647	.444
11.0000	.529	.444
12.5000	.412	.333
13.5000	.235	.222
14.5000	.176	.222
17.0000	.118	.222
21.0000	.118	.111
24.5000	.059	.111
32.5000	.059	.000
40.0000	.000	.000

The test result variable(s): LCI5diff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

TUG TI1



Diagonal segments are produced by ties.

Area Under the Curve
.346

Coordinates of the Curve

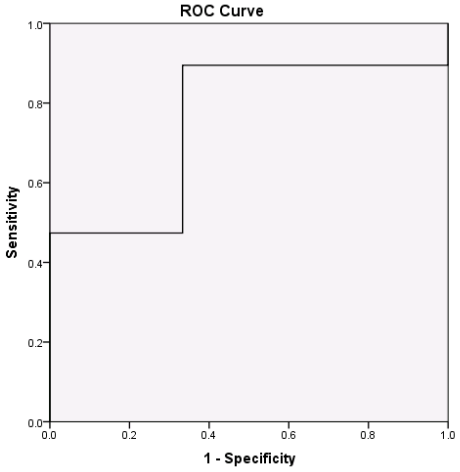
Test Result Variable(s): TUGdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-22.7000	1.000	1.000
-20.1000	.944	1.000
-17.6800	.889	1.000
-16.5250	.833	1.000
-15.7000	.778	1.000
-14.6100	.778	.889
-13.9200	.722	.889
-13.1400	.667	.889
-12.3100	.667	.778
-12.0850	.611	.778
-11.6650	.556	.778
-10.7400	.500	.778
-9.2950	.444	.778
-8.4200	.389	.778
-8.2300	.389	.667
-7.2550	.389	.556
-6.2000	.333	.556
-5.6150	.333	.444
-5.2750	.278	.444
-5.0500	.278	.333
-4.8000	.222	.333
-4.2800	.167	.333

-2.5650	.167	.222
-1.2150	.111	.222
-.3150	.111	.111
1.2200	.056	.111
6.8400	.000	.111
12.7000	.000	.000

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

TUG TI2



Area Under the Curve
.754

Coordinates of the Curve

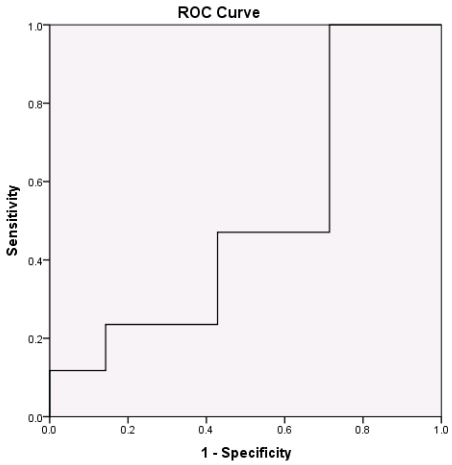
Test Result Variable(s): TUGdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-20.1700	1.000	1.000
-19.0850	.947	1.000
-18.4850	.895	1.000
-15.7000	.895	.667
-12.3850	.895	.333
-9.6450	.842	.333
-7.9200	.789	.333
-7.7400	.737	.333
-7.2500	.684	.333
-6.6100	.632	.333
-6.1600	.579	.333
-5.7200	.526	.333

-4.9700	.474	.333
-4.2300	.474	.000
-3.4700	.421	.000
-2.4850	.368	.000
-1.6600	.316	.000
-1.3050	.263	.000
-.7700	.211	.000
-.0450	.158	.000
.4900	.105	.000
6.2600	.053	.000
12.7200	.000	.000

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

TUG TI3



Area Under the Curve
.504

Coordinates of the Curve

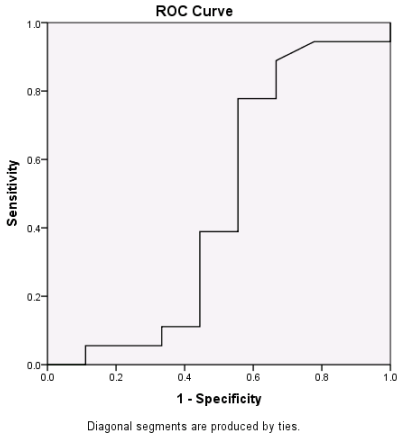
Test Result Variable(s): TUGdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-36.8600	1.000	1.000
-31.9900	1.000	.857
-25.5400	1.000	.714
-22.0900	.941	.714
-20.2900	.882	.714
-18.8500	.824	.714

-18.3250	.765	.714
-17.3500	.706	.714
-15.7750	.647	.714
-14.9350	.588	.714
-14.3750	.529	.714
-13.6250	.471	.714
-13.1850	.471	.571
-12.7950	.471	.429
-12.2200	.412	.429
-10.4700	.353	.429
-8.1100	.294	.429
-7.0500	.235	.429
-6.4250	.235	.286
-5.5050	.235	.143
-3.7400	.176	.143
-2.1200	.118	.143
-1.4450	.118	.000
2.6400	.059	.000
7.4400	.000	.000

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

2MWT TI1



Area Under the Curve
.466

Coordinates of the Curve

Test Result Variable(s): Walktestdiff

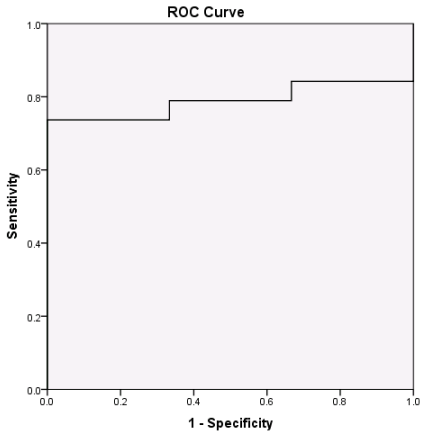
Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-20.1000	1.000	1.000
-17.3000	.944	1.000
-8.3500	.944	.889
-.6000	.944	.778
1.3750	.889	.667
3.7250	.833	.667
4.7500	.778	.667
5.5250	.778	.556
6.5000	.722	.556
6.7750	.667	.556
7.2500	.611	.556
9.3000	.556	.556
11.6750	.500	.556
12.6500	.444	.556
13.4500	.389	.556
15.0750	.389	.444
16.8500	.333	.444
19.0500	.278	.444
21.2750	.222	.444
22.1250	.167	.444
25.7500	.111	.444
31.2000	.111	.333
37.3750	.056	.333

47.3750	.056	.222
57.9000	.056	.111
65.1500	.000	.111
68.6000	.000	.000

The test result variable(s): Walktestdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

2MWT TI2



Area Under the Curve
.789

Coordinates of the Curve

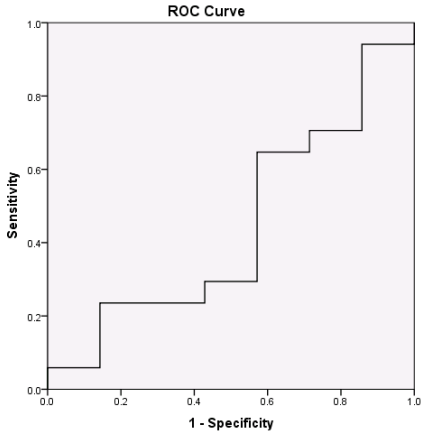
Test Result Variable(s): Walktestdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-39.0500	1.000	1.000
-23.3500	.947	1.000
-8.2000	.895	1.000
-5.3750	.842	1.000
-1.2500	.842	.667
.8500	.789	.667
1.9500	.789	.333
4.4000	.737	.333
8.2000	.737	.000
10.8250	.684	.000
11.4250	.632	.000
14.3250	.579	.000

18.1250	.526	.000
19.7500	.474	.000
20.7500	.421	.000
22.6000	.368	.000
25.8500	.316	.000
29.2250	.263	.000
31.9750	.211	.000
41.1750	.105	.000
58.5750	.053	.000
68.9000	.000	.000

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

2MWT TI3



Area Under the Curve
.445

Coordinates of the Curve

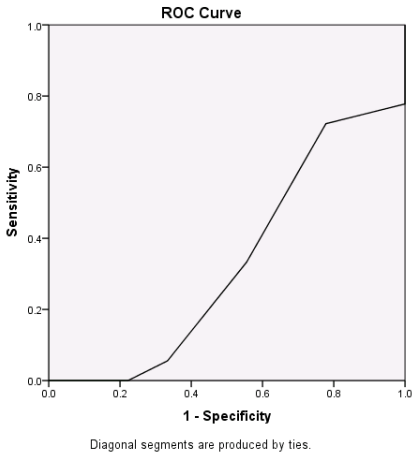
Test Result Variable(s): Walktestdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-32.3000	1.000	1.000
-24.9000	.941	1.000
-12.2000	.941	.857
-.4500	.882	.857
7.0500	.824	.857
10.3000	.765	.857
12.2000	.706	.857

17.5750	.706	.714
22.4000	.647	.714
26.0750	.647	.571
31.2250	.588	.571
33.5750	.529	.571
35.7500	.471	.571
37.5500	.412	.571
39.3500	.353	.571
41.0500	.294	.571
41.4750	.294	.429
46.5000	.235	.429
53.1500	.235	.286
55.3750	.235	.143
56.8500	.176	.143
66.2250	.118	.143
76.5500	.059	.143
107.0250	.059	.000
136.5000	.000	.000

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

SCS TI1



Diagonal segments are produced by ties.

Area Under the Curve
.330

Coordinates of the Curve

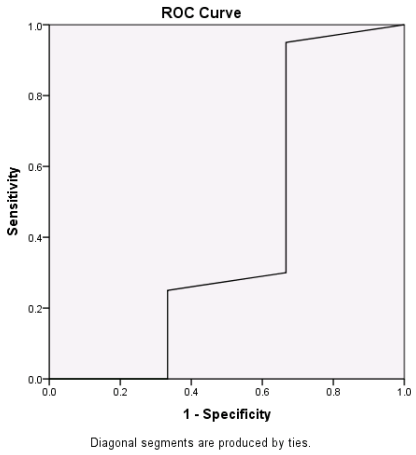
Test Result Variable(s): SCSdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-5.0000	1.000	1.000
-3.5000	.889	1.000
-2.5000	.833	1.000
-1.5000	.778	1.000
-.5000	.722	.778
.5000	.333	.556
2.0000	.056	.333
3.5000	.000	.222
4.5000	.000	.111
6.0000	.000	.000

The test result variable(s): SCSdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

SCS TI2



Area Under the Curve
.417

Coordinates of the Curve

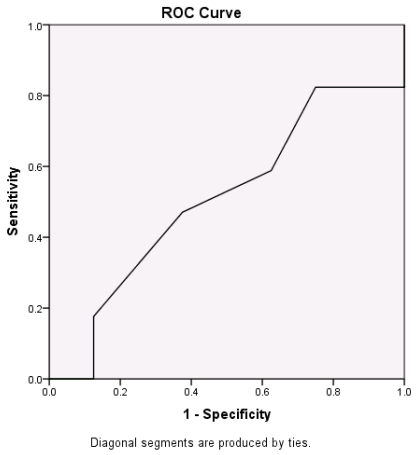
Test Result Variable(s): SCSdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-4.0000	1.000	1.000
-2.5000	.950	.667
-1.5000	.900	.667
-.5000	.750	.667
.5000	.300	.667
1.5000	.250	.333
2.5000	.200	.333
3.5000	.100	.333
5.0000	.000	.333
7.0000	.000	.000

The test result variable(s): SCSdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

SCS TI3



Area Under the Curve
.507

Coordinates of the Curve

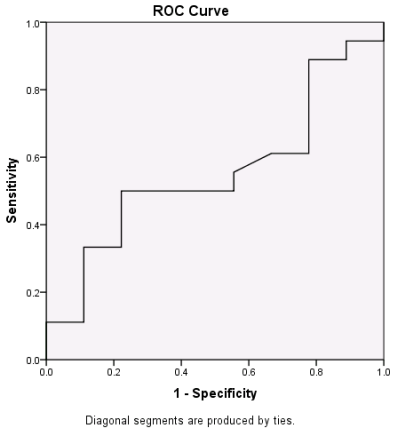
Test Result Variable(s): SCSdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-5.0000	1.000	1.000
-3.5000	.941	1.000
-2.0000	.824	1.000
-.5000	.824	.750
.5000	.588	.625
1.5000	.471	.375
2.5000	.176	.125
3.5000	.059	.125
4.5000	.000	.125
6.0000	.000	.000

The test result variable(s): SCSdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

EQ-5D-index TI1



Diagonal segments are produced by ties.

Area Under the Curve
.552

Coordinates of the Curve

Test Result Variable(s): EQ5Dindexdiff

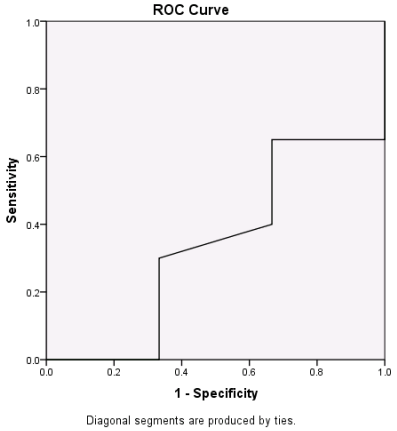
Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-1.2330	1.000	1.000
-.1510	.944	1.000
-.0625	.944	.889
-.0480	.889	.889
-.0350	.889	.778
-.0150	.833	.778
.0085	.667	.778
.0245	.611	.778
.0365	.611	.667
.0440	.556	.556
.0520	.500	.556
.0635	.500	.444
.0725	.500	.333
.0780	.500	.222
.0835	.444	.222
.0890	.389	.222
.0980	.333	.222
.1080	.333	.111
.1280	.278	.111
.1970	.222	.111
.2530	.167	.111
.3310	.111	.111
.4320	.111	.000

.4815	.056	.000
1.5050	.000	.000

The test result variable(s): EQ5Dindexdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

EQ-5D-index TI2



Area Under the Curve
.333

Coordinates of the Curve

Test Result Variable(s): EQ5Dindexdiff

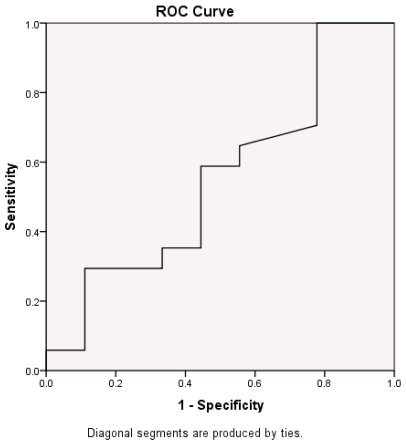
Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-1.4700	1.000	1.000
-.4205	.950	1.000
-.3590	.900	1.000
-.2375	.850	1.000
-.1185	.800	1.000
-.0905	.750	1.000
-.0695	.700	1.000
-.0540	.650	1.000
-.0215	.650	.667
-.0010	.600	.667
.0005	.500	.667
.0070	.450	.667
.0265	.400	.667

.0410	.300	.333
.0430	.250	.333
.0690	.200	.333
.0960	.150	.333
.0990	.100	.333
.1225	.050	.333
.1985	.000	.333
1.2520	.000	.000

The test result variable(s): EQ5Dindexdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

EQ-5D-index TI3



Area Under the Curve
.549

Coordinates of the Curve

Test Result Variable(s): EQ5Dindexdiff

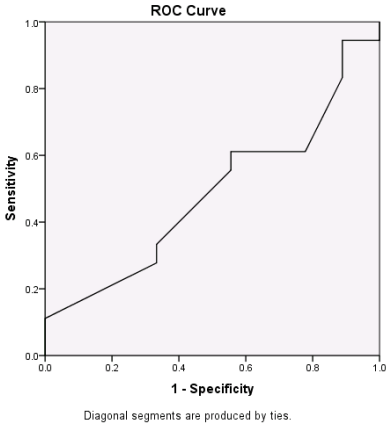
Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-1.7030	1.000	1.000
-.5865	1.000	.889
-.2990	1.000	.778
-.1215	.882	.778
-.0720	.824	.778
-.0240	.765	.778
-.0095	.706	.778

.0095	.647	.556
.0245	.588	.556
.0315	.588	.444
.0370	.529	.444
.0450	.471	.444
.0610	.412	.444
.0915	.353	.444
.1105	.353	.333
.1130	.294	.333
.1245	.294	.222
.1680	.294	.111
.2465	.235	.111
.3120	.176	.111
.3415	.118	.111
.4000	.059	.111
.5245	.059	.000
1.5990	.000	.000

The test result variable(s): EQ5Dindexdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

EQ-5D-VAS TI1



Area Under the Curve
.485

Coordinates of the Curve

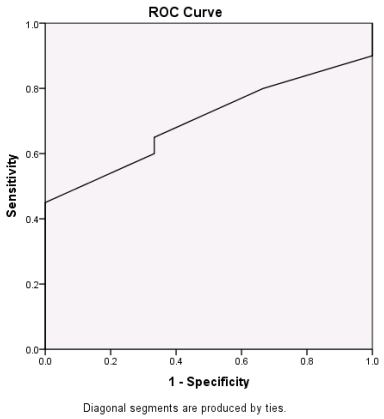
Test Result Variable(s): EQ5DVASdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-36.0000	1.000	1.000
-25.0000	.944	1.000
-10.0000	.944	.889
-2.5000	.833	.889
1.0000	.611	.778
2.5000	.611	.667
3.5000	.611	.556
4.5000	.556	.556
6.5000	.333	.333
9.0000	.278	.333
12.5000	.167	.111
17.5000	.111	.000
21.0000	.000	.000

The test result variable(s): EQ5DVASdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

EQ-5D-VAS TI2



Diagonal segments are produced by ties.

Area Under the Curve
.700

Coordinates of the Curve

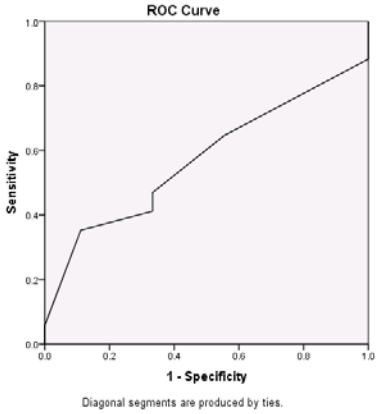
Test Result Variable(s): EQ5DVASdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-21.0000	1.000	1.000
-17.5000	.900	1.000
-12.5000	.800	.667
-9.0000	.650	.333
-6.5000	.600	.333
-4.5000	.450	.000
-2.5000	.400	.000
-.5000	.350	.000
2.5000	.250	.000
6.5000	.150	.000
11.5000	.100	.000
22.5000	.050	.000
31.0000	.000	.000

The test result variable(s): EQ5DVASdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

EQ-5D-VAS TI2



Area Under the Curve
.572

Coordinates of the Curve

Test Result Variable(s): EQ5DVASdiff

Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
-21.0000	1.000	1.000
-17.5000	.941	1.000
-12.5000	.882	1.000
-7.5000	.824	.889
-2.5000	.647	.556
1.0000	.471	.333
3.5000	.412	.333
7.5000	.353	.111
17.5000	.059	.000
26.0000	.000	.000

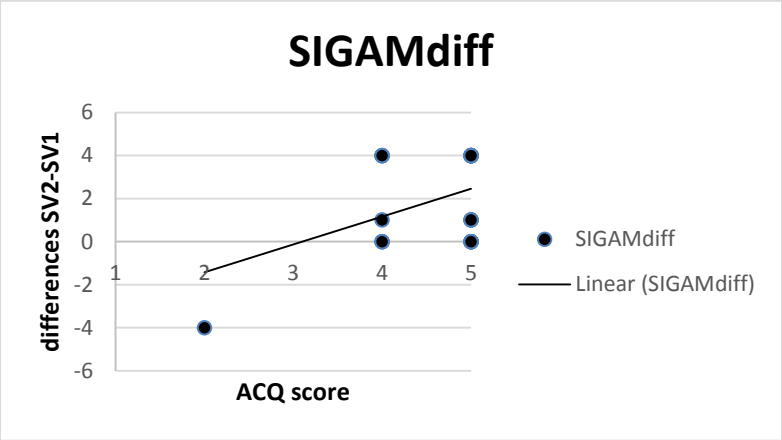
The test result variable(s): EQ5DVASdiff has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

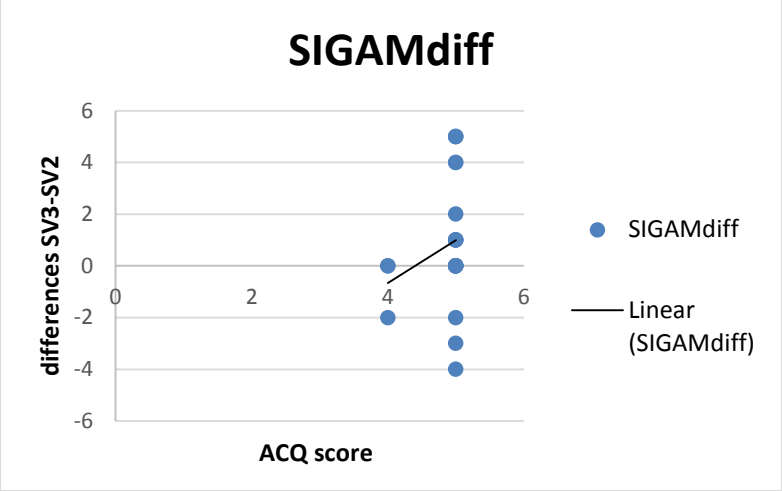
Appendix 15

Correlation Plots: ACQ vs differences

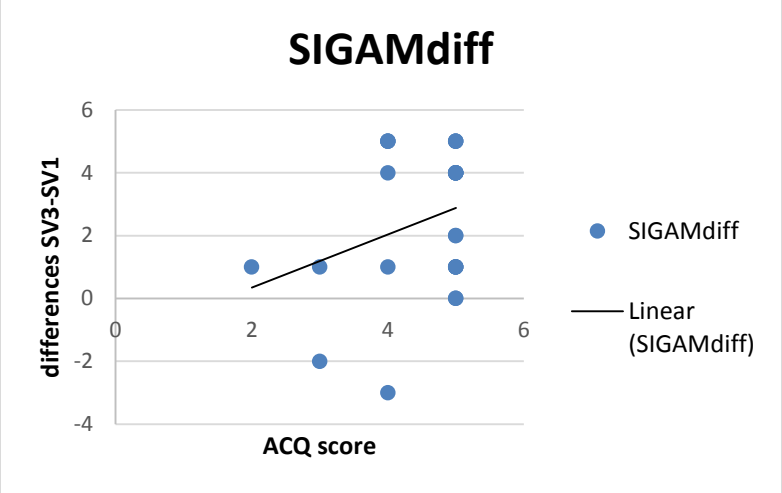
Time Interval 1 (SV2-SV1)



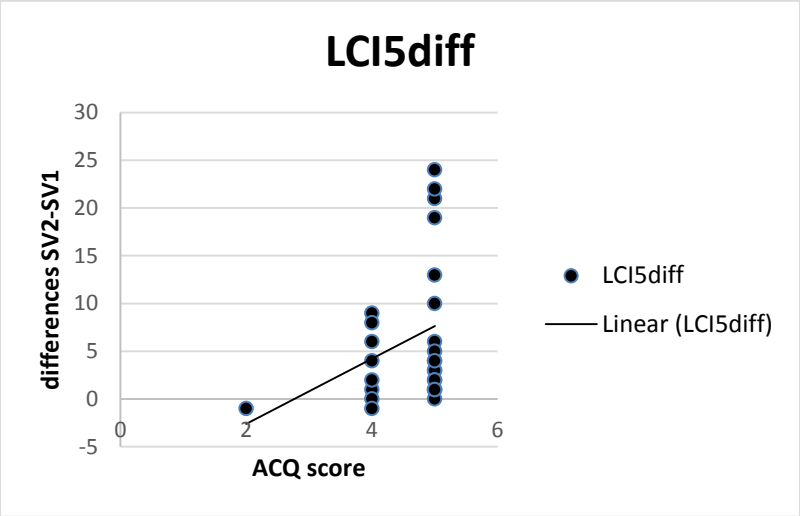
Time Interval 2 (SV3 – SV2)



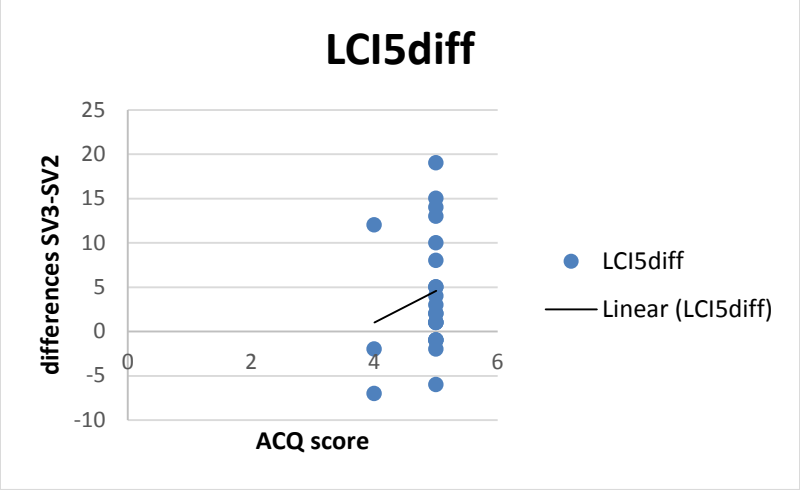
Time Interval SV3-SV1



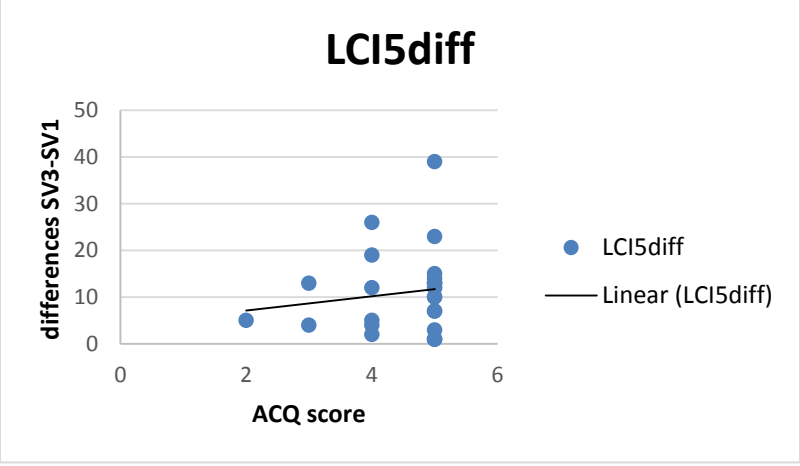
Time Interval 1 (SV2-SV1)



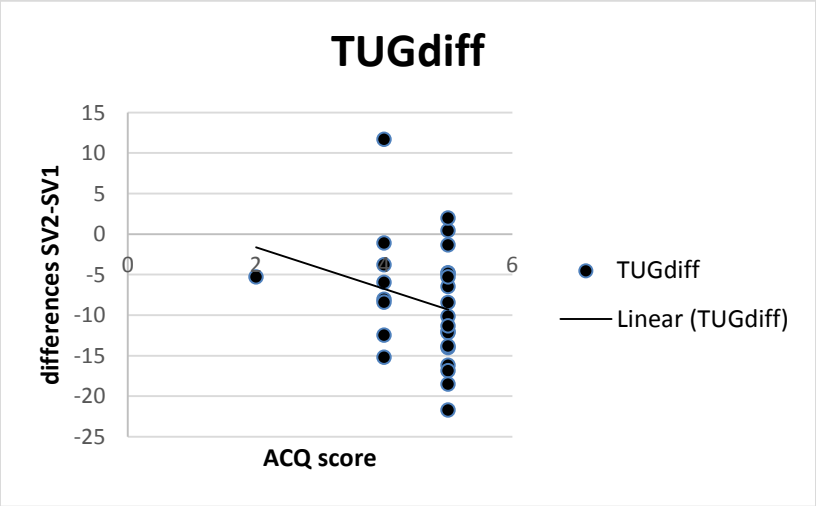
Time Interval 2 (SV3 – SV2)



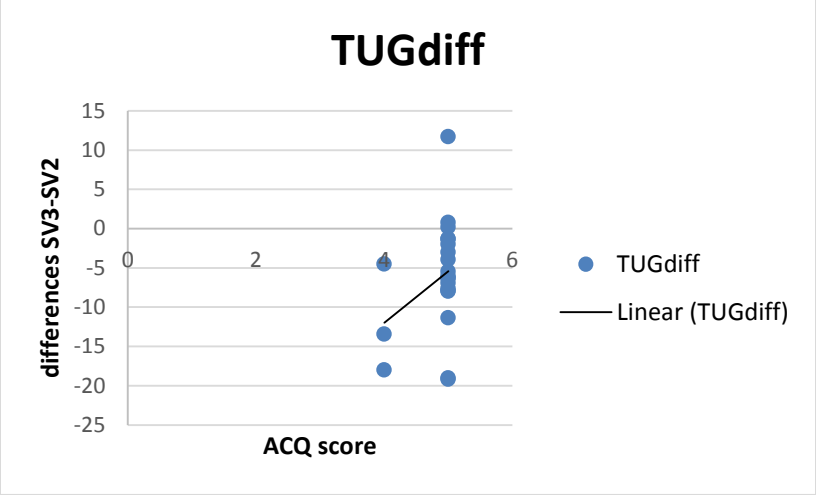
Time Interval SV3-SV1



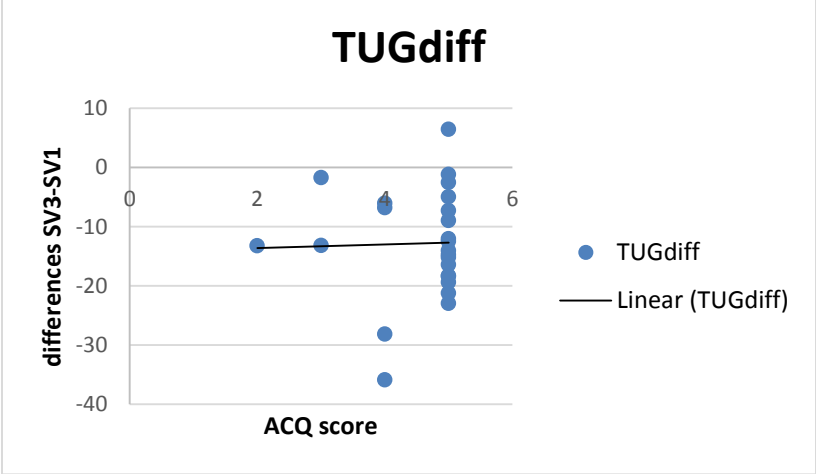
Time Interval 1 (SV2-SV1)



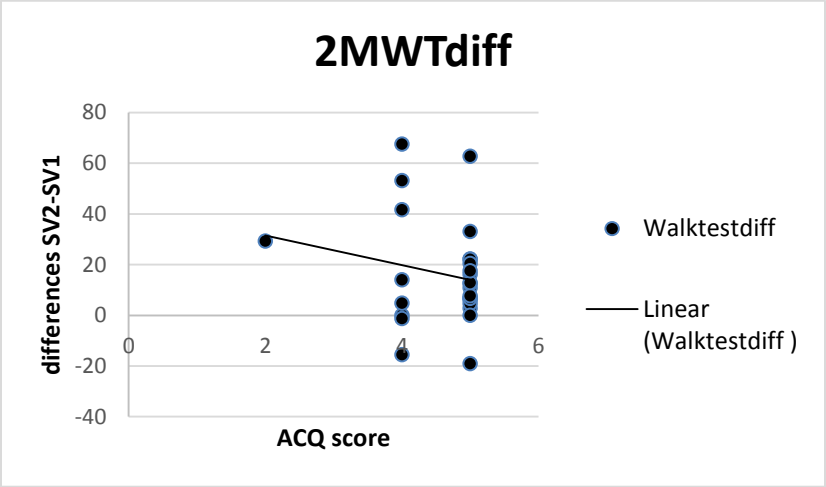
Time Interval 2 (SV3 – SV2)



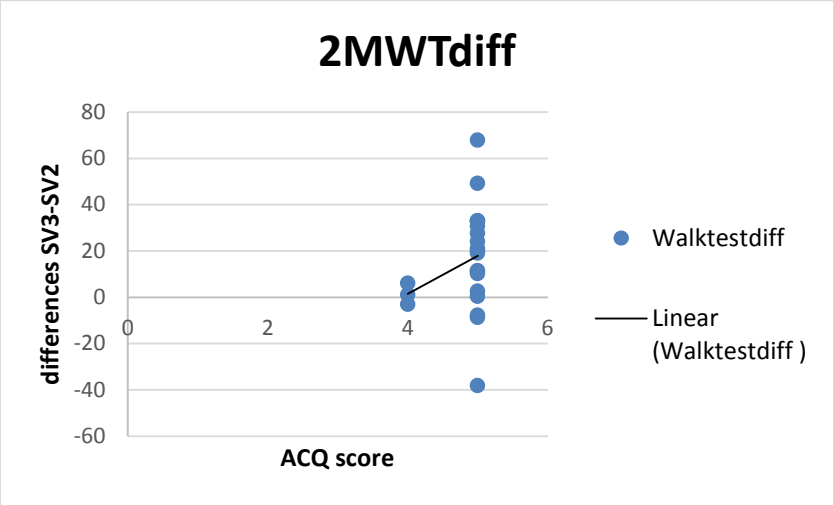
Time Interval SV3-SV1



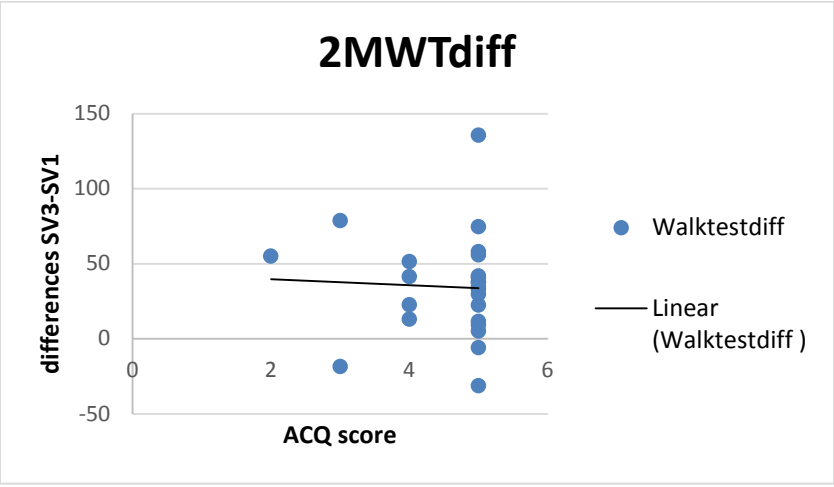
Time Interval 1 (SV2-SV1)



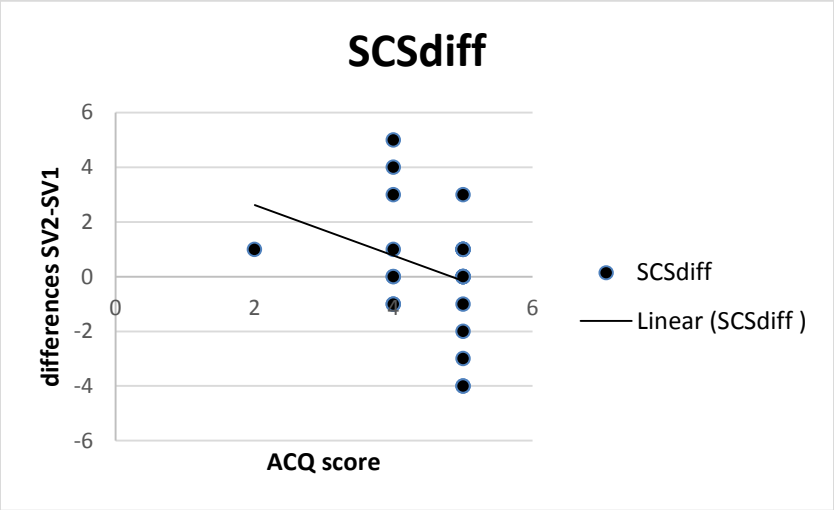
Time Interval 2 (SV3 – SV2)



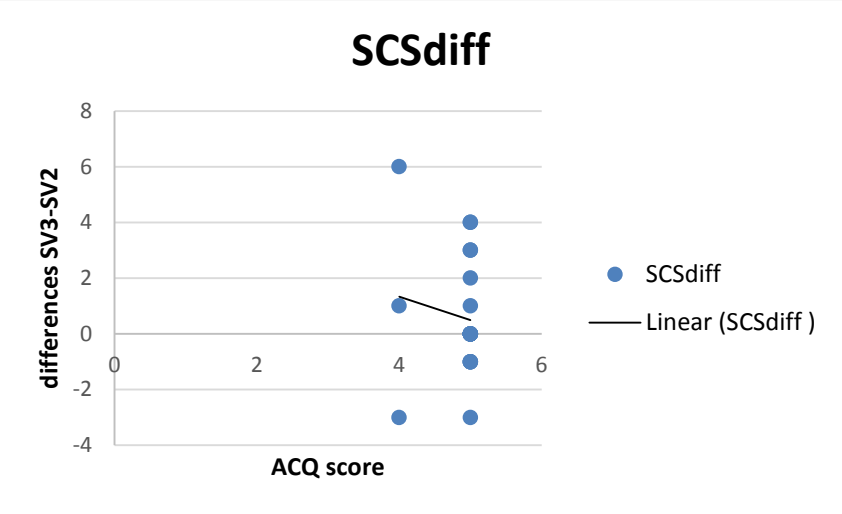
Time Interval SV3-SV1



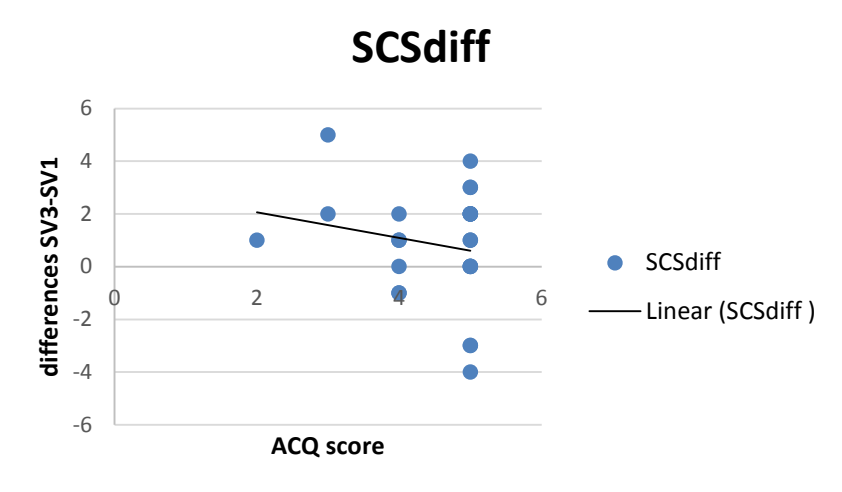
Time Interval 1 (SV2-SV1)



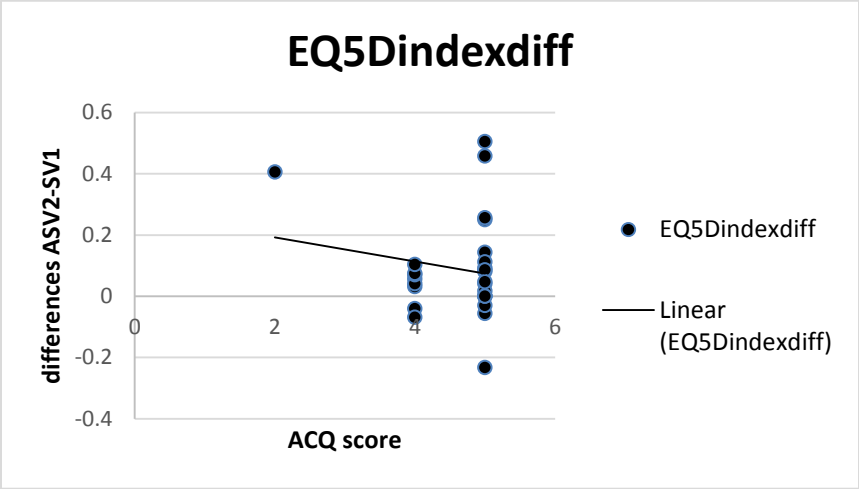
Time Interval 2 (SV3 – SV2)



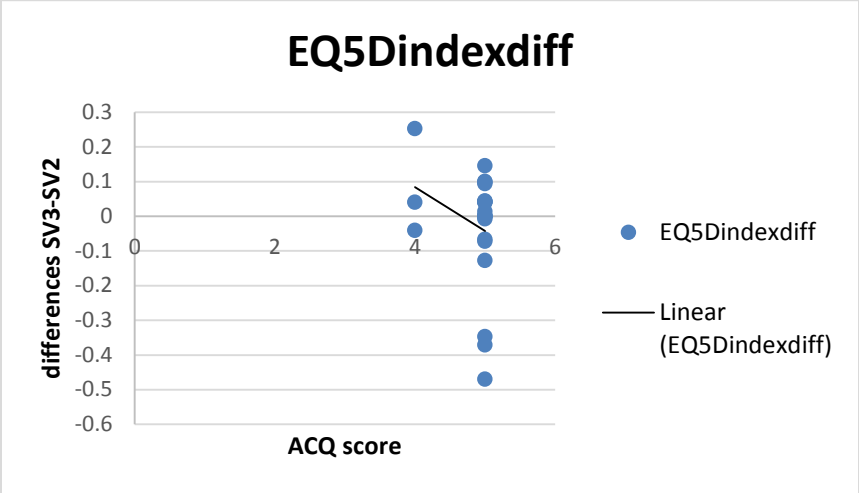
Time Interval SV3-SV1



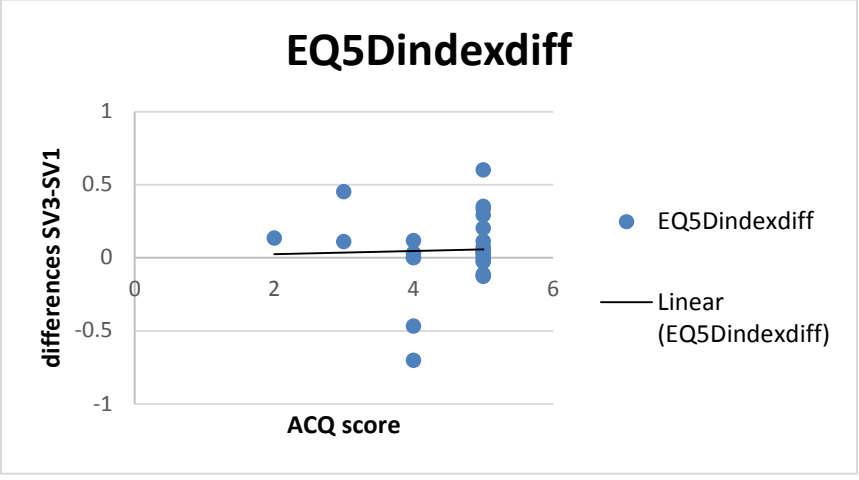
Time Interval 1 (SV2-SV1)



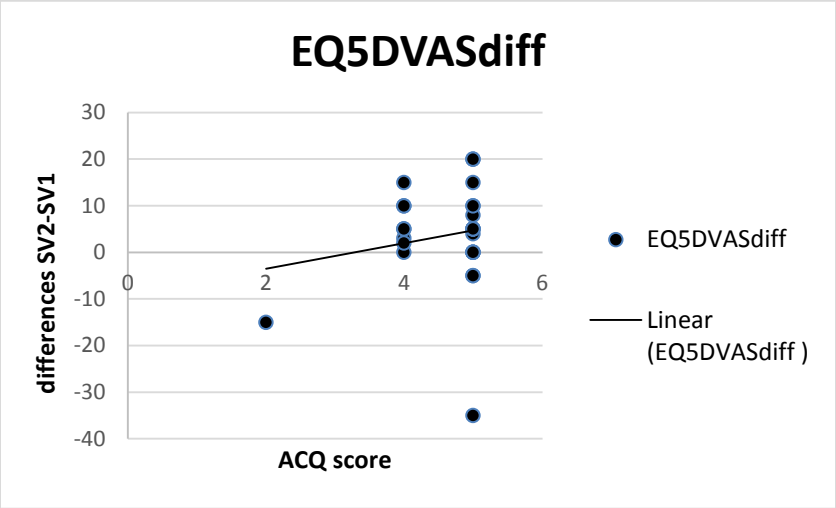
Time Interval 2 (SV3 – SV2)



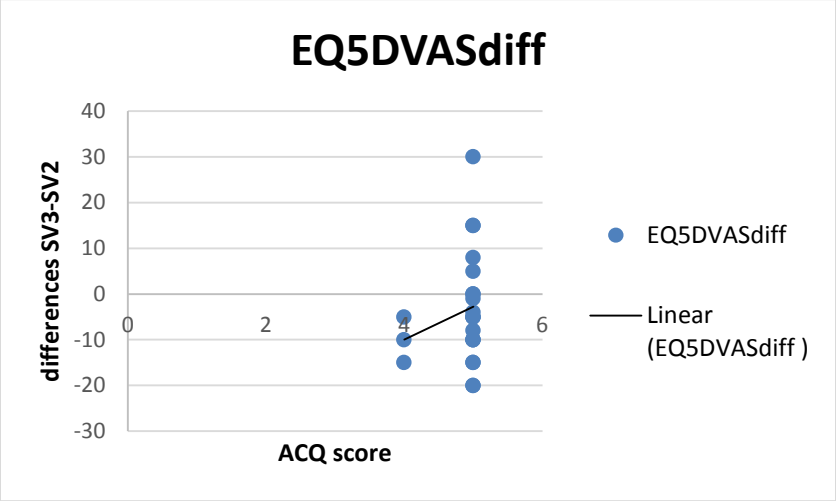
Time Interval SV3-SV1



Time Interval 1 (SV2-SV1)



Time Interval 2 (SV3 – SV2)



Time Interval SV3-SV1